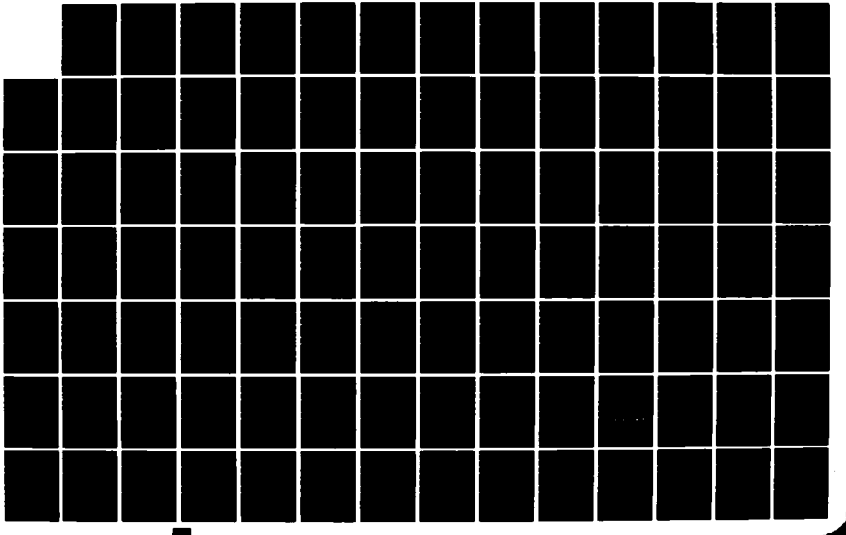
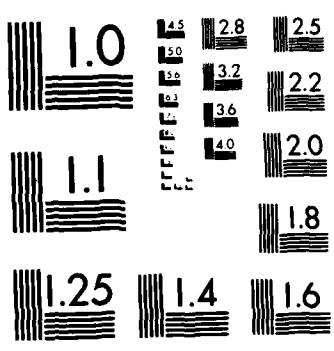
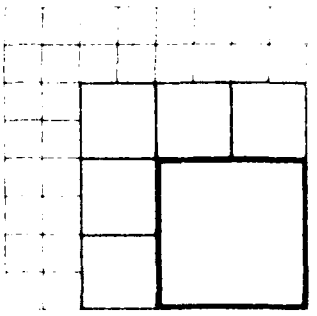
AD-A153 605 TOWARDS A STATISTICAL ANALYSIS OF GENETIC SEQUENCES DATA WITH PARTICULAR.. (U) MASSACHUSETTS INST OF TECH CAMBRIDGE STATISTICS CENTER S P ARSENIS MAR 85 1/2

UNCLASSIFIED TR-36-ONR N00014-74-C-0555 F/G 6/3 NL







STATISTICS CENTER

Massachusetts Institute of Technology

77 Massachusetts Avenue Rm. E40-111, Cambridge, Massachusetts 02139 (617) 253-8722

(2)

TOWARDS A STATISTICAL ANALYSIS OF GENETIC SEQUENCES DATA WITH PARTICULAR REFERENCE TO PROTEIN SEQUENCES

BY

SPYROS P. ARSENIS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

TECHNICAL REPORT NO. ONR 36

MARCH 1985

PREPARED UNDER CONTRACT

N00014-74-C-0555 (NR-609-001)

FOR THE OFFICE OF NAVAL RESEARCH

TOWARDS A STATISTICAL ANALYSIS OF GENETIC SEQUENCES DATA WITH PARTICULAR REFERENCE TO PROTEIN SEQUENCES.

by

SPYROS P. ARSENIS

ABSTRACT

This report develops a variety of character matrices as graphical tools for the visual examination of genetic sequences and in particular protein sequences. The NNC, PNC, BNC1, BNC2 and BNC3 matrices are designed to filter noise without severely suppressing signals in the CC matrix. The Matrix Smear of a character matrix is introduced as a measure of signals and noise in the matrix. The asymptotic distribution of the smears of the CC and NNC matrices are derived under the independence model. The asymptotic result is used in conjunction with exact confidence intervals from diagonal smears to automate partially the visual examination of character matrices. A generalized likelihood ratio procedure is developed to automate fully the detection of signals in two protein sequences. A simulation study has proven the procedure to be powerful and robust in detecting signals of success probability .90 and length 9 implanted within noisy binary strings of length 291 characters and success probability .15.

Some Key Words: Genetic sequences, DNA, Matrix Smear, Character Matrix Graphics

AMS 1980 subject classification. Primary 62P10

TABLE OF CONTENTS

1. INTRODUCTION, BIOLOGICAL BACKGROUND AND NOMENCLATURE.

The subject of this research is the development of a statistical methodology to analyze protein and DNA sequence data. Various data analytic tools presented here; their development was motivated from the examination of fourteen DNA sequences which encode proteins forming the eggshell of the American silkmoth Antherea polyphemus. The genes were sequenced in the laboratory of professor Fotis Kafatos.

The question that was initially posed by Fotis Kafatos, was to cluster the fourteen genes on the basis of their similarities within regions where similarities had already been detected. A measure of similarity between strings was developed and its application to the regions where the genes had been detected to be similar produced clusters that made good biological sense.

To find out if there were other regions where the fourteen genes were similar, graphical ways to represent the data were required. Through these it became clear that the genes shared similarities far more extensive than previously detected and that there was a lot of structure within each gene, basically in the form of consecutive repeats of a basic repeat unit.

Chapter 2 presents a variety of character matrices as graphical tools to allow the investigator to look into string data. These matrices are designed so as to reduce the matrix smear - which is a measure of "signals" and "noise" in the data - without suppressing "signals". Chapter 3 presents an aymptotic result for the distribution of the smear of some of the matrices of chapter 2, under the assumption that strings

are written independently between and within themselves. Chapter 4 compares the matrix to the diagonal smears to "automate" the visual examination of character matrices. Chapter 5 develops a machine examination of character matrices by listing the significant substrings of the words which maximize a generalized log-likelihood ratio for the hypothesis that for two parameters p_0 and p_1, $p_0 \leq p_1$, the probability of a match is smaller than p_0 vs. the alternative hypothesis that it is larger than p_1. Chapter 1 now presents the biological background necessary to pose questions relating to genetic sequence data and concludes with the presentation of the chorion data set. The compendium is based on Dayhoff [6], Hood [8], Mahan [10], and Watson [15].

Observed via a microscope, chromosomes are paired threadlike structures in the nuclei of living cells. Since the beginning of this century, chromosomes were recognized to be responsible for the transmission of the hereditary properties of organisms via their subunits, called genes. As little had been known about their structure at the molecular level, however, genes were considered as black boxes until rather recently.

A chromosome is a giant DNA molecule. Proposed by Watson and Crick in 1953, the structure of DNA is that of two intertwined strands giving the molecule the shape of a double helix as illustrated in figure 1-1.

The backbone of each strand is provided by the sugar molecule deoxyribose. The structural formula of deoxyribose is shown in figure 1-2. On the one apex of the pentagonal ring stands an oxygen (O) atom, the other four being occupied by carbon (C) atoms. On the deoxyribose molecule there are five C atoms indexed by the integers 1, 2, 3, 4, and

5 in figure 1-2. Attached to the 1 C atom is one of the four molecules: adenine (A), guanine (G), cytosine (C) and thymine (T). These four together with uracil (U), which will be referred to later as a building block of DNA, are called bases.

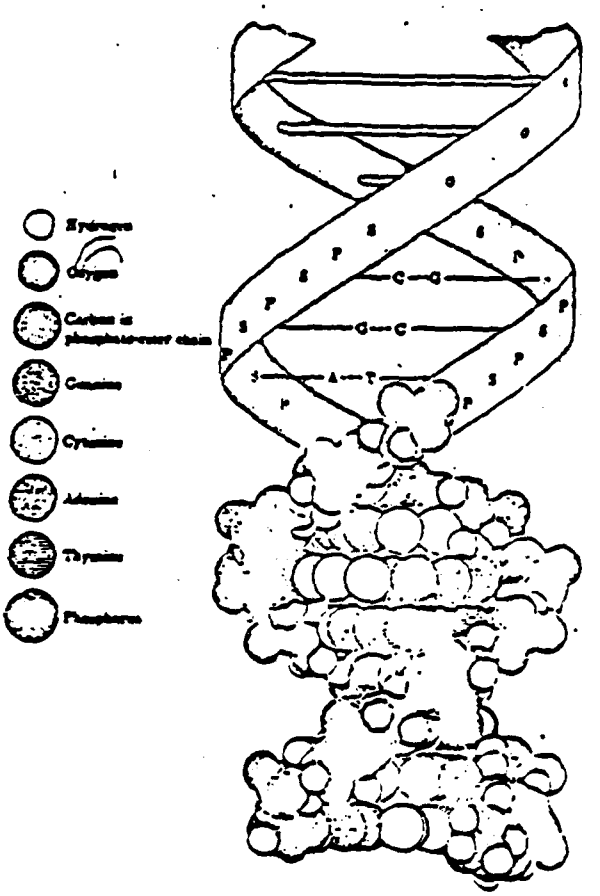


Figure 1-1. The structrure of the DNA molecule.

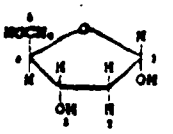


Figure 1-2. The structure of the deoxyribose molecule.

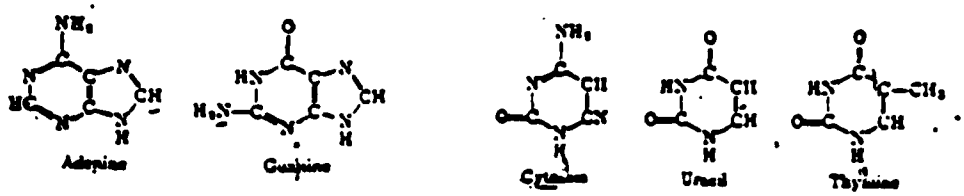


Figure 1-3. The structural formulae of the five bases adenine, guanine, cytosine, uracil and thymine.

The structural formulae of the five bases are shown in figure 1-3. To the 3 and 5 C atom sites of deoxyribose are attached phosphate groups (PO_4^{---}) that provide the links between successive sugar molecules in the DNA strand as illustrated in figure 1-4.

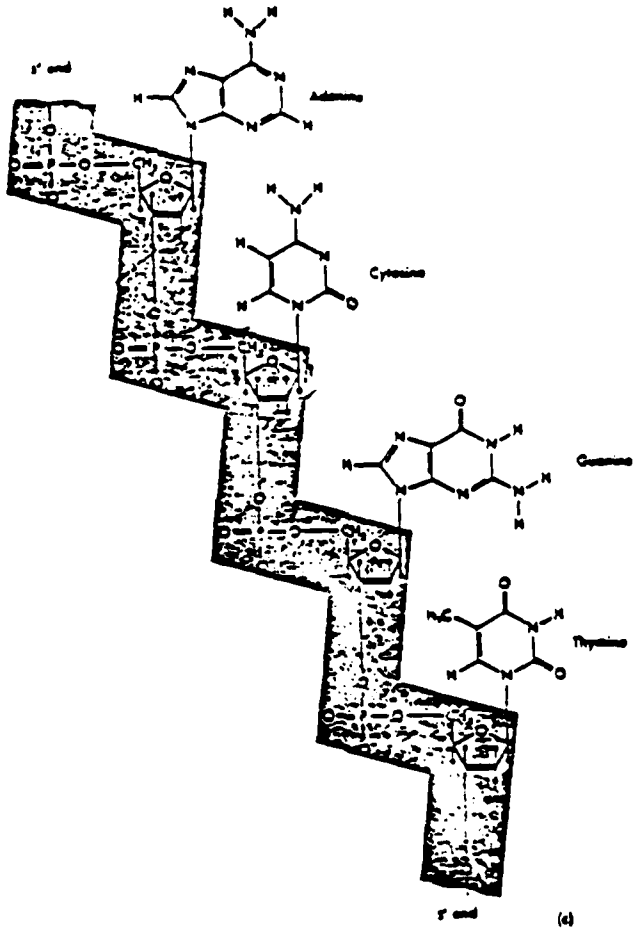


Figure 1-4. The structure of a strand of a DNA molecule.

The combination of the deoxyribose molecule with one of the bases and the phosphate group is called a nucleotide. The phosphate and the deoxyribose always being the same, nucleotides are denoted by the base molecules T, C, A, G, or U which are attached to the deoxyribose.

The helical structure of DNA is made possible by bonds among bases in opposite strands. In particular, thymines bind to adenines and guanines to cytosines ("base pairing rules"). Consequently, DNA may be presented by the sequence of nucleotides in one strand, together with the direction in which that sequence is read. The convention established in the biochemical literature is that a sequence of letters from the alphabet of T, C, A, G represents the nucleotides from the chain end on the 5 C atom of deoxyribose to that on the 3 C site. With this convention, DNA sequence data will be considered as words written in the alphabet of the four bases {T,C,A,G}. They will be denoted as finite sequences $X = (X_1,\ldots,X_n)$, for $X_i \varepsilon \{T,C,A,G\}$.

At the molecular level, a gene is a piece of the DNA molecule usually several hundred base letters long. A gene encodes and, under certain conditions, directs the synthesis of a protein as is sketched later on in this section. The protein coding portion of a gene starts with the letters ATG and ends with one of TAA, TAG, or TGA.

Proteins are molecules found throughout living organisms acting as enzymes (catalyzing various biochemical reactions) or forming membranes of cells and other cellular structures (playing a structural role). The building blocks of proteins are the amino acids. Table 1-1 gives the alphabet in which the twenty amino acids are conventionally abbreviated.

Table 1-1. 1-letter abbreviations for the twenty amino acids.

1	Phenylalanine	F	11	Isoleucine	I
2	Leucine	L	12	Methionine	M
3	Serine	S	13	Threonine	T
4	Tyrosine	Y	14	Asparagine	N
5	Cysteine	C	15	Lysine	K
6	Tryptophan	W	16	Valine	V
7	Proline	P	17	Alanine	A
8	Histidine	H	18	Aspartic	D
9	Glutamine	Q	19	Glutamic	E
10	Arginine	R	20	Glycine	G

For our purposes, and in the absence of other information about their structure, proteins are words written in the alphabet of the twenty letters of table 1-1 and denoted as finite sequences $X = (X_1,\ldots,X_n)$, for all X_i in the alphabet of the twenty letters. A protein sequence is written in the direction in which its encoding DNA sequence is conventionally written, each amino acid encoded by three consecutive nucleotides as will be explained below. Proteins and DNA sequences will be interchangeably referred to as words or strings; stretches of the above will be referred to as syllables or substrings.

The synthesis of a protein is directed by its corresponding gene through the following two step mechanism:

(1) Transcription of DNA to mRNA. One of the two strands of the DNA molecule acts as a template which appropriate enzymes copy into RNA, a chemically similar molecule. RNA is a single stranded molecule built up of nucleotides bound to each other as in DNA. The bases in the RNA nucleotides are A, G, C, and U. They are respectively copied from the

T, C, G and A bases of the DNA strand under transcription. The transcribed RNA strand subsequently undergoes "splicing". In particular, regions of the RNA strands, called "introns" for intervening, are removed and the remaining regions, called "exons", are joined together to form the messenger RNA. (mRNA)

(2) Translation of the mRNA to the protein. The mRNA acts as a template which in conjunction with other components of the cell (ribosomes, tRNA, etc) directs the assembly of a string of corresponding amino acids as specified by the genetic code.

The genetic code is shown in table 1-2. It maps each triplet of consecutive nucleotides, called a codon, to an amino acid except for codons UAA, UAG, UGA. The latter codons monitor the end of the protein coding region of the gene and are called terminator codons. Codon AUG is used as an initiator or for encoding methionine internal to the protein chain. Since 61 codons are mapped into 20 amino acids, amino acids are bound to be encoded by more than one codon.

Table 1-2. The genetic code with codons entered in a three way table.

UUU	F	UCU	S	UAU	Y	UGU	C
UUC	F	UCC	S	UAC	Y	UGC	C
UUA	L	UCA	S	UAA	Term	UGA	Term
UUG	L	UCG	S	UAG	Term	UGG	W
CUU	L	CCU	P	CAU	H	CCU	R
CUC	L	CCC	P	CAC	H	CGC	R
CUA	L	CCA	P	CAA	Q	CGA	R
CUG	L	CCG	P	CAG	Q	CGG	R
AUU	I	ACU	T	AAU	N	AGU	S
AUC	I	ACC	T	AAC	N	AGC	S
AUA	I	ACA	T	AAA	K	AGA	R
AUG	M	ACG	T	AAG	K	AGG	R
GUU	V	GCU	A	GAU	D	GGU	G
GUC	V	GCC	A	GAC	D	GGC	G
GUA	V	GCA	A	GAA	E	GGA	G
GUG	V	GCG	A	GAA	E	GGG	G

Supported by fossil and biochemical evidence, the fundamental evolutionary scenario of biology, postulates that billions of years ago, life on earth existed in a simple ancestral form. While organisms evolved from their common ancestor, numerous mutations accumulated on their genetic material. Mutations occur in individual organisms by chance; over time they may spread through or disappear from the population. Their laws are studied in evolutionary biology and population genetics and are not directly relevant in the present discussion.

The fundamental scenario adapts to the biochemical level of description of organisms as follows: living organisms undergo mutations on their genetic material. Mutations of two kinds have been observed. A base may substitute another in a DNA strand and give rise to a point mutation. Fractions of a gene, whole genes, or microscopically visible pieces of chromosomes may duplicate, become deleted, or translocate. Segmental mutations refer to the above events incurring on fractions of genes. Mutations are said to be selectively deleterious to the individual organism on which they are imposed if they increase the likelihood of that the individual organism dies or leaves fewer descendants. Other mutations may offer the organism selective advantages, or may be selectively neutral. Gravely deleterious mutations are censored by natural selection; selectively neutral or even slightly deleterious mutations may survive or even become fixed in the population by chance.

Figure 1-5 presents the coding portions of the fourteen genes under analysis. The genes are given as 292, 292a, 292b, 609, 18, 18b, 18c, 401, 401a, 401b, 408, 10, 10a, and 10b. On the basis of their extensive similarities, genes 292, 292a, and 292b, are collectively called 292 copies. (Similarly for 18, 18b, and 18c etc.) The first seven

genes will be referred to as family A. Family B comprises the last seven genes.

On the basis of when their protein products are formed during the formation of the eggshell, 292 copies and gene 509 form the middle A subfamily; the copies of 18 form the late A subfamily. The late B subfamily is made of the copies of 401 while gene 408 and the copies of 10 form the middle B subfamily.

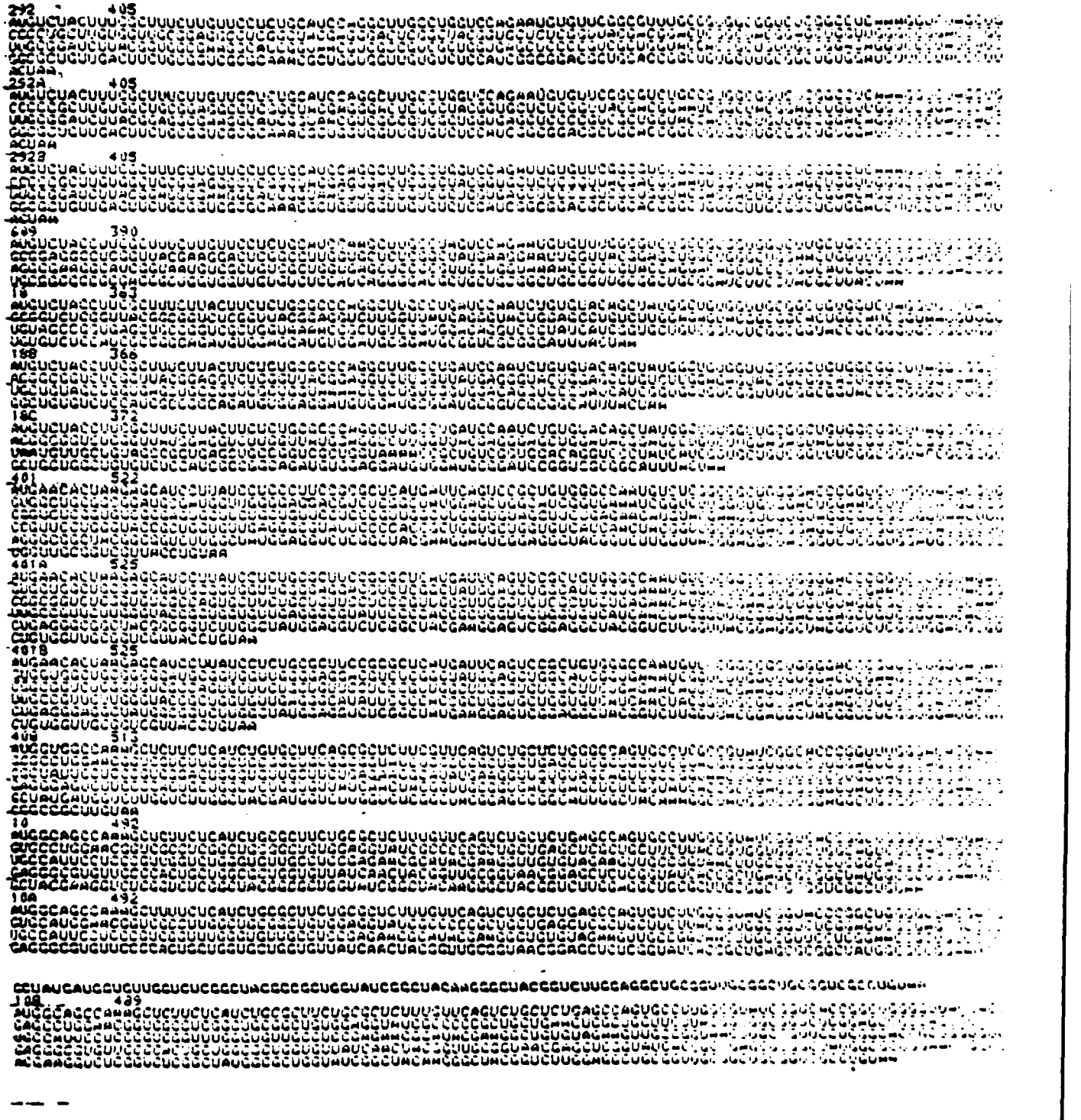


Figure 1-5. The DNA sequences for the coding regions of the chorion genes of families A and B.

2. CHARACTER MATRICES AS EXPLORATORY TOOLS FOR GENETIC SEQUENCES DATA

The proteins encoded by the chorion genes under analysis are listed in figure 2-1. Human vision is inadequate to detect structure within or similarities between the proteins as they are presented. This chapter introduces a variety of character matrices which proved useful in bringing out similarities between different proteins and repeats within proteins. Character matrices have been constructed for DNA and amino acid sequences throughout this research. In this chapter they are introduced in the general context of two words and illustrated for some of the proteins of figure 2-1.

Let $\underline{X} = (X_1,\ldots,X_m)$ and $\underline{Y} = (Y_1,\ldots,Y_n)$ be words written in the alphabet $\{a_1,\ldots,a_s\}$. $\underline{X}$ and $\underline{Y}$ may or may not be the same. The Crude Character (CC) matrix for $\underline{X}$ and $\underline{Y}$ is defined by:

$$M_{i,j} = \begin{cases} X_i & \text{if } X_i=Y_j \\ \text{" " (blank)} & \text{otherwise.} \end{cases} \qquad (2.1)$$

$i=1,\ldots,m$ and $j=1,\ldots,n$.

The idea of using two dimensional arrays to look into string data appeared first, latently, in figure 1 of Needleman and Wunch [11] in their exposition of an algorithm to compute the longest common subsequnce between two words. CC matrices were also explicitly constructed in Gibbs and McIntyre [7]

Character matrices are useful exploratory tool for looking into sequence data because a substring common to the two words shows up as a diagonal in the CC matrix for the words. Figure 2-2 presents the CC matrix for the proteins encoded by genes 292 and 18B. Two major, relatively solid diagonals can be distinguished on the CC matrix of

figure 2-2. The longest diagonal consists of entries $(M_{75,63},\ldots,M_{129,117})$ and indicates that syllables $(X_{75},\ldots,X_{129})$ and $(Y_{63},\ldots,Y_{117})$ are similar in the sense that $X_i=Y_{i+12}$ for $i=63,\ldots,117$ except for a few occasional mismatches. The structure of the matrix block corresponding to $(X_{40},\ldots,X_{51})$ and $(Y_{29},\ldots,Y_{50})$ will become clear in matrices to be presented later. For the moment it is noted that the longest diagonal in the block is $(M_{43,33},\ldots,M_{59,49})$ and parallel to it and within the block run other shorter diagonals. In a character matrix for two words, the appearance of parallel diagonals at a substring of one of the words signifies the existence of internal repeats in the other word, as illustrated in figure 2-3.

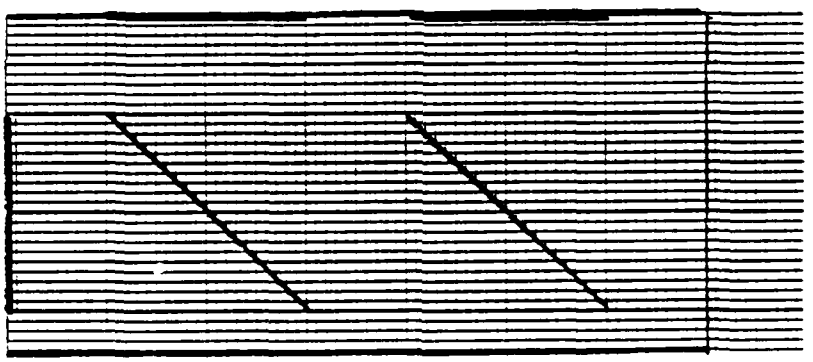


Figure 2-3. Parallel diagonals at a substring of $\underline{X}$ are due to and signify repeats of the substing in $\underline{Y}$.

The CC matrix for $\underline{X}=\underline{Y}$ brings out internal repeats within word $\underline{X}$. It is symmetric and its entries M_{ii} are nonblank. The CC matrix for protein 292 is presented in figure 2-4. (It is not square only because the characters of the LP used are rectangular.) The diagonal string $(M_{42,52},\ldots,M_{51,61})$ marked on figure 2-4, runs parallel to the solid diagonal of the matrix and is formed by the repeat of the syllable

$(X_{42},\ldots,X_{51})$ as $(X_{52},\ldots,X_{61})$ except for one mismatch.

The usefulness of CC matrices is limited by two factors: their size and "noise" associated with them.

A common line printer can print up to 8 lines per inch vertically and up to 132 characters per line horizontally. Therefore, when printed on a line printer, the CC matrix for a word of 500 characters (a length common for DNA sequence data) is longer than six feet. To diminish the size of the matrices the investigator has to prepare successive photoreductions at the expense of papercutting and paperpasting. This limitation may also be circumvented by presenting character matrices on a plotter. A digital plotter applies a large grid (for example of 4095 by 3124 sites) on a sheet of paper of desirable dimensions. A character matrix in blanks and dots may then be plotted by placing dots instead of alphabet characters at the appropriate grid sites.

The second limitation of CC matrices is more serious. In attempting to trace diagonals the human eye is distracted by characters which are are bound to appear only because of the composition of the words. In particular, if the counts of alphabet characters $a_1,\ldots,a_s$ in $\underline{X}$ and $\underline{Y}$ are $m_1,\ldots,m_s$ and $n_1,\ldots n_s$ respectively, the CC matrix for $\underline{X}$ and $\underline{Y}$ contains $\sum_1^s m_i n_i$ nonblank characters. Hence, the ratio of nonblank characters to all the characters in the matrix is:

$$S(\underline{X},\underline{Y}) = \sum_1^s \frac{m_i}{m}\frac{n_i}{n}, \qquad (2\text{-}2)$$

where m_i/m and n_i/n are the relative frequencies of a_i in the two words. $S(\underline{X},\underline{Y})$ in (2-2) will be called the <u>matrix smear for the CC</u>

matrix of $\underline{X}$ and $\underline{Y}$.

If $\underline{X}$ is independent of $\underline{Y}$ and $\{X_t\}$ t=1,...m and $\{Y_t\}$ t=1,...,n are I.I.D. with $Pr(X_t=a_i)=p_i$ and $Pr(Y_t=a_i)=q_i$ for all t and i, the matrix smear is a sample estimator of the parameter

$$\sigma = \sum_{1}^{s} p_i q_i. \qquad (2\text{-}3)$$

σ will be called the theoretical smear for the CC matrix of $\underline{X}$ and $\underline{Y}$. Under the above independent assumptions the theoretical smear is the probability of a nonblank character in the matrix.

The matrix smear of the CC matrix for two different words ranges from 0 (for words with no alphabet character in common) to 1 (for words written in one letter). The matrix smear for the CC matrix for the word $\underline{X}$ is

$$S(\underline{X}) = \sum_{1}^{s} \left(\frac{m_i}{m} \right)^2. \qquad (2\text{-}4)$$

$S(\underline{X})$ is minimized when $m_1=\ldots=m_s$. The minimum attained is the inverse of size of the alphabet in which $\underline{X}$ is written. Table 2-1a lists to the second decimal digit the smears for all pairs of chorion proteins.

Table 2-1a. Smears of CC matrices for all pairs of chorion proteins.

	292	292A	292B	609	18	18B	18C	401	401A	401B	408	10	10A	10B
292	.12													
292A	.12	.12												
292B	.12	.12	.12											
609	.12	.12	.12	.12										
18	.13	.13	.13	.13	.15									
18b	.13	.13	.13	.13	.15	.15								
18C	.13	.13	.13	.13	.15	.15	.15							
401	.13	.13	.13	.13	.15	.15	.15	.15						
401A	.13	.13	.13	.13	.15	.15	.15	.15	.15					
401B	.13	.13	.13	.13	.15	.15	.15	.15	.15	.15				
408	.12	.12	.12	.12	.13	.13	.14	.13	.13	.13	.13			
10	.12	.12	.12	.12	.13	.13	.13	.13	.13	.13	.12	.12		
10A	.12	.12	.12	.12	.13	.13	.13	.13	.13	.13	.12	.12	.12	
10B	.12	.12	.12	.12	.13	.13	.13	.13	.13	.13	.12	.12	.12	.12

Smears for all pairs range from 12% to 15%. Matrix smears within subfamilies are stable as can be seen from table 2-1b below.

Table 2-1b. Range of smears of CC matrices within subfamilies of Chorion proteins.

	Middle A	Late A	Late B	Middle B
Middle A	.12			
Late A	.13	.15		
Late B	.13	.15	.15	
Middle B	.12	.13-.14	.13	.12

The matrix smear specifies the number of non blank characters appearing in a given matrix and can be thought of as a measure of "signal" and "noise" in the data. Is it possible to reduce the smear without substantially supressing diagonals in the matrix? Recall that the (i,j)th entry of the CC matrix was defined by comparing X_i to Y_j. Now consider the Next Neighbour Considered (NNC) character matrix for $\underline{X}$ and $\underline{Y}$, defined as:

$$M_{i,j} = \begin{cases} X_i & \text{if } X_i=Y_j \text{ and } X_{i+1}=Y_{j+1} \\ " \ " & \text{otherwise.} \end{cases} \qquad (2\text{-}5)$$

$i=1,\ldots,m-1$ and $j=1,\ldots,n-1$.

Figure 2-5 presents the NNC matrix for proteins 292 and 18B. It clarifies the extensive repeat structure in the block formed by $(X_{40},\ldots,X_{61})$ and $(X_{29},\ldots,X_{50})$ and brings out the features that 292 and 18B share in common. If the syllable $(a_i a_j)$ occurs $m_{i,j}$ and $n_{i,j}$ times in $\underline{X}$ and $\underline{Y}$, then

$$\sum_{i=1}^{s}\sum_{j=1}^{s} m_{ij} = m-1, \qquad \sum_{i=1}^{s}\sum_{j=1}^{s} n_{ij} = n-1,$$

and the ratio of nonblank entries of the NNC matrix to the total number of matrix entries is:

$$S(\underline{X},\underline{Y}) = \sum_{1}^{s}\sum_{1}^{s} \frac{m_{i,j}}{m-1}\frac{n_{i,j}}{n-1}. \qquad (2\text{-}6)$$

The ratio of equation (2-6) will be called <u>the smear of the NNC matrix</u> of $\underline{X}$ and $\underline{Y}$. The smear of the NNC matrix for one word becomes

$$S(\underline{X}) = \sum_{1}^{s}\sum_{1}^{s} \left(\frac{m_{i,j}}{m-1}\right)^2$$

and attains a minimum equal to the inverse of the square of the alphabet size. Table 2-2 lists up to the second decimal digit the smears of the NNC matrices for all chorion proteins.

Table 2-2. Smears of NNC matrices for all pairs of chorion proteins.

	292	292A	292B	609	18	18B	18C	401	401A	401B	408	10	10A	10B
292	.02													
292A	.02	.02												
292B	.02	.02	.02											
609	.02	.02	.02	.02										
18	.02	.02	.02	.02	.03									
18B	.02	.02	.02	.02	.03	.03								
18C	.02	.02	.02	.02	.03	.03	.03							
401	.02	.02	.02	.02	.02	.02	.02	.03						
401A	.02	.02	.02	.02	.02	.02	.02	.03	.03					
401B	.02	.02	.02	.02	.02	.02	.02	.03	.03	.03				
408	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02			
10	.02	.01	.01	.01	.02	.02	.02	.02	.02	.02	.02	.02		
10A	.01	.01	.01	.01	.02	.02	.02	.02	.02	.02	.02	.02	.02	
10B	.02	.01	.01	.01	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02

Smears of the NNC matrices range from 1% to 3%. Those for the NNC matrices for the same protein vary between 2% and 3% compared to the minimum .25%. For proteins 292 and 18B the smear of 13% for the CC matrix is reduced to 2% for the NNC matrix.

NNC matrices eliminate a number of nonblank characters appearing on CC matrices that only blur diagonal strings. On the other hand, corresponding to two syllables that are identical except for one mismatch, the CC matrix produces a diagonal that is broken at one point

while the NNC matrix breaks the diagonal at two entries. This suggests a third character matrix for which $M_{i,j}$ are defined after comparing the 3-letter syllables (X_{i-1},X_i,X_{i+1}) and (Y_{j-1},Y_j,Y_{j+1}) as follows:

$$M_{i,j} = \begin{cases} X_i & \text{if } X_i=Y_j \text{ and } (X_{i-1}=Y_{j-1} \text{ or } X_{i+1}=Y_{j+1}) \\ * & \text{if } X_{i-1}=Y_{j-1},\ X_i\neq Y_j,\ X_{i+1}=Y_{j+1} \\ \text{' '} & \text{if otherwise.} \end{cases} \qquad (2\text{-}7)$$

$i=2,\ldots,m-1$ and $j=2,\ldots n-1$.

The matrix defined in equation (2-7) will be called Both Neighbours Considered and abbreviated by BNC1, the index 1 appended to the acronym BNC to distinguish it from other matrices defined by comparing 3-letter syllables. Figure 2-6 presents the BNC1 matrix for proteins 292 and 18B. The BNC1 matrix allows up to nonconsecutive mismatches in similar strings without breaking their diagonal. Table 2-3 presents the smears of the BNC1 matrices for all chorion proteins.

Table 2-3. Smears of BNC1 matrices for all pairs of chorion proteins.

	292	292A	292B	609	18	18B	18C	401	401A	401B	408	10	10A	10B
292	.04													
292A	.04	.04												
292B	.04	.04	.04											
609	.04	.04	.04	.05										
18	.05	.05	.05	.05	.06									
18B	.05	.05	.05	.05	.06	.06								
18C	.05	.05	.05	.05	.06	.06	.06							
401	.05	.05	.05	.05	.06	.06	.07	.07						
401A	.05	.05	.05	.05	.06	.06	.07	.07	.07					
401B	.05	.05	.05	.05	.06	.06	.07	.07	.07	.07				
408	.05	.04	.04	.05	.06	.06	.06	.06	.06	.06	.06			
10	.04	.04	.04	.04	.05	.05	.06	.06	.06	.06	.05	.05		
10A	.04	.04	.04	.04	.05	.05	.06	.06	.06	.06	.05	.05	.05	
10B	.04	.04	.04	.04	.05	.05	.05	.06	.06	.06	.05	.05	.05	.05

The table indicates that the smears of the BNC1 matrices for chorion proteins range from 4% to 7%. The smear of the BNC1 matrix for proteins 292 and 18B is calculated to be 5%, between that of the CC (13%) and the NNC matrix (2%).

Another BNC matrix, called BNC2, is defined by

$$M_{i,j} = \begin{cases} X_i & \text{if } X_i=Y_j \text{ and } (X_{i-1}=Y_{j-1} \text{ or } X_{i+1}=Y_{j+1}) \\ \text{" "} & \text{otherwise.} \end{cases} \qquad (2\text{-}8)$$

$i=2,\ldots,m-1$ and $j=2,\ldots n-1$.

The BNC2 differs from the BNC1 matrix in that it suppresses the "*" of equation (2-7). For comparison purposes, the BNC2 matrix for proteins 292 and 18B is presented in figure 2-7.

Finally we define the BNC3 matrix as:

$$M_{i,j} = \begin{cases} X_i & \text{if } X_{i-1}=Y_{j-1}\ ,\ X_i=Y_j \text{ and } X_{i+1}=Y_{j+1} \\ \text{" "} & \text{if otherwise.} \end{cases} \qquad (2\text{-}9)$$

$i=2,\ldots,m-1$ and $j=2,\ldots n-1$.

The BNC3 matrix for proteins 292 and 18B is presented in figure 2-8. Table 2-4 lists the smears for the BNC3 matrices for all pairs of chorion proteins up to the second decimal digit. Smears less than .01 are not entered in the table.

Table 2-4. Smears of BNC3 matrices for all pairs of chorion proteins.

	292	292A	292B	609	18	18B	18C	401	401A	401B	408	10	10A	10B
292	.01													
292A	.01	.01												
292B	.01	.01	.01											
609	.01	.01	.01	.01										
18	.01	.01	.01	.01	.01									
18B	.01	.01	.01	.01	.01	.02								
18C	.01	.01	.01	.01	.01	.01	.02							
401					.01	.01	.01	.01						
401A	.01				.01	.01	.01	.01	.01					
401B	.01				.01	.01	.01	.01	.01	.01				
408					.01	.01	.01	.01	.01	.01	.01			
10					.01	.01	.01	.01	.01	.01	.01	.01		
10A							.01	.01	.01	.01	.01	.01	.01	
10B					.01	.01	.01	.01	.01	.01	.01	.01	.01	.01

As can be seen from table 2-4 the smears of the BNC3 matrices for chorion proteins range up to 2%. Those for the same protein vary from 1%

to 2% compared to the minimum $1/20^3$ = .0125%. The BNC3 matrix "filters" the data rather severely and suppresses diagonals that were discernible in less restrictive matrices presented previously.

The entries of all matrices defined so far are blanks, asterisks or alphabet characters. It is clear that in a quantitative assessment of diagonals the types of matches and mismatches should be taken into consideration, matches of rare letters being more "significant" than those between frequent letters. However, visual examinations of character matrices are not elaborate enough to take the nature of matches or mismatches into account. In whichever matrix is available, the investigator is searching for long diagonals with a large number of matching nonblank characters relative to the length of the diagonal. Thus for purposes of visual examination a matrix entry may be reduced to a blank or a non-blank character.

The five types of character matrices introduced in this chapter are conceptually and mathematically related. The (i,j)th entry of the NNC matrix was defined after comparing X_i to Y_j and their next (right) neighbours X_{i+1} and Y_{j+1}. Instead one might compare the previous (left) neighbours X_{i-1} and Y_{j-1} and construct the Previous Neighbour Considered (PNC) matrix. The superposition of the PNC to the NNC produces the BNC2 matrix.

The design of various character matrices to reduce the smear and enable the investigator to discern existing diagonals, was previouly called "filtering" of the data. The term has not only a heuristic appeal; for the NNC, PNC, and BNC3 matrices it is used appropriately in a technical sense too. Indeed, we can consider these character matrices as CC types of matrices on the data after they are transformed

appropriately. In perticular, consider transforming the sequences $\{X_t\}$, $t=1,\ldots,m$ and $\{Y_t\}$, $t=1,\ldots n$ as:

$$\tilde{\underline{X}}_t = \begin{bmatrix} X_t \\ X_{t+1} \end{bmatrix} \qquad \tilde{\underline{Y}}_s = \begin{bmatrix} Y_s \\ Y_{s+1} \end{bmatrix} \qquad \begin{matrix} t=1,\ldots,m-1 \\ s=1,\ldots,n-1 \end{matrix}$$

Then the NNC matrix can be thought of as a CC type of matrix on the transformed data. The transformations corresponding to the PNC and BNC3 matrices are

$$\tilde{\underline{X}}_t = \begin{bmatrix} X_t \\ X_{t-1} \end{bmatrix} \qquad \tilde{\underline{Y}}_s = \begin{bmatrix} Y_s \\ Y_{s-1} \end{bmatrix} \qquad \begin{matrix} t=2,\ldots,m \\ s=2,\ldots,n \end{matrix}$$

and

$$\tilde{\underline{X}}_t = \begin{bmatrix} X_{t-1} \\ X_t \\ X_{t+1} \end{bmatrix} \qquad \tilde{\underline{Y}}_s = \begin{bmatrix} Y_{s-1} \\ Y_s \\ Y_{s+1} \end{bmatrix} \qquad \begin{matrix} t=2,\ldots,m-1 \\ s=2,\ldots,n-1 \end{matrix}$$

respectively.

The NNC, PNC, BNC1, BNC2 and BNC3 character matrices were designed in order to reduce the noise in the CC matrices and make signals easily discernible. Of those, the NNC and PNC and BNC3 matrices suppress signals as well. A syllable of length L present in common in $\underline{X}$ and $\underline{Y}$ gives rise to a diagonal string of length L-1 for the NNC and PNC matrices and L-2 for the BNC3 matrix. The BNC2 matrix does not suppress signals but does not allow for mismatches; when a substring is common to the two words except for a mismatch, the diagonal corresponding to the syllable carries a blank character at the site of the mismatch. While filtering noise, the BNC1 may be thought of as enhancing signals as it does not allow occasional nonconsecutive mismatches in a syllable which is otherwise shared by $\underline{X}$ and $\underline{Y}$, to brake the diagonal corresponding to it.

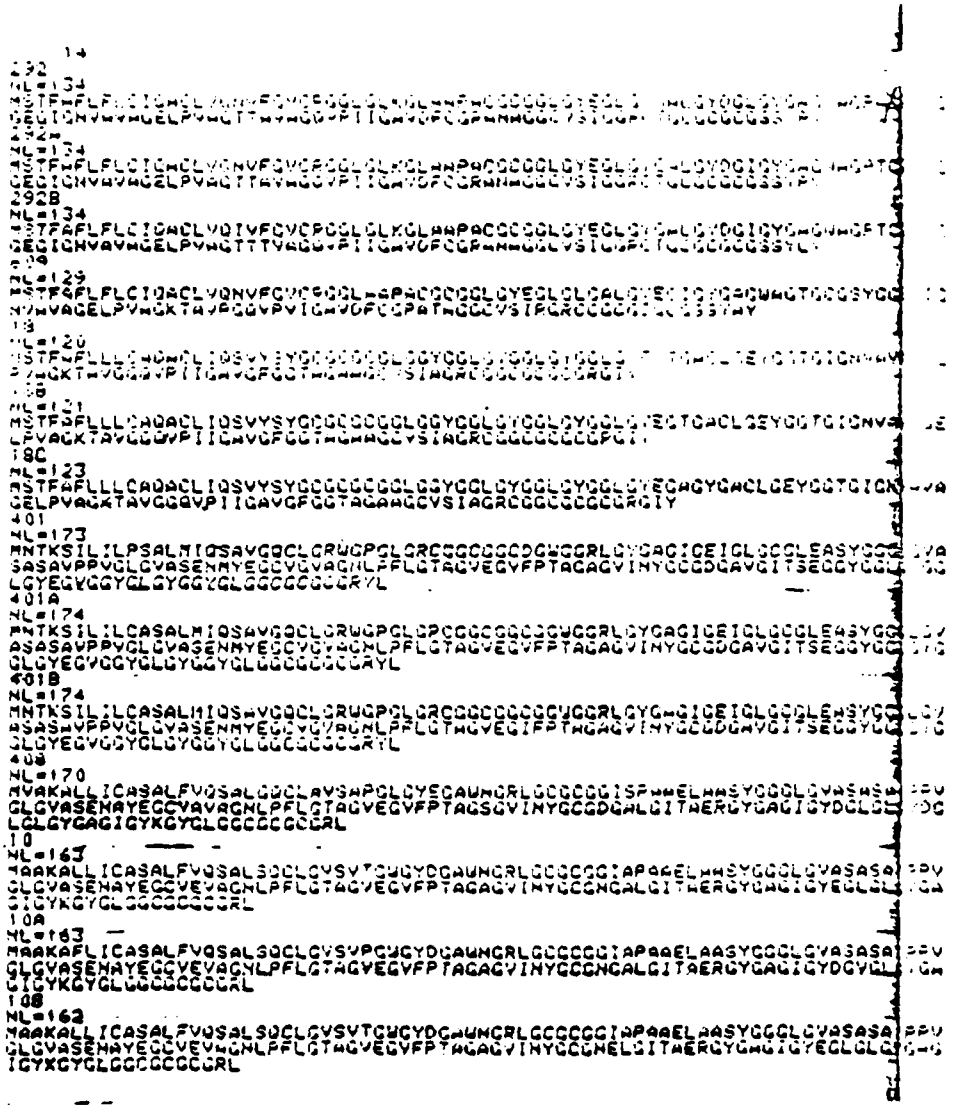


Fig. 2-1. Proteins encoded by the fourteen chorion genes under analysis.

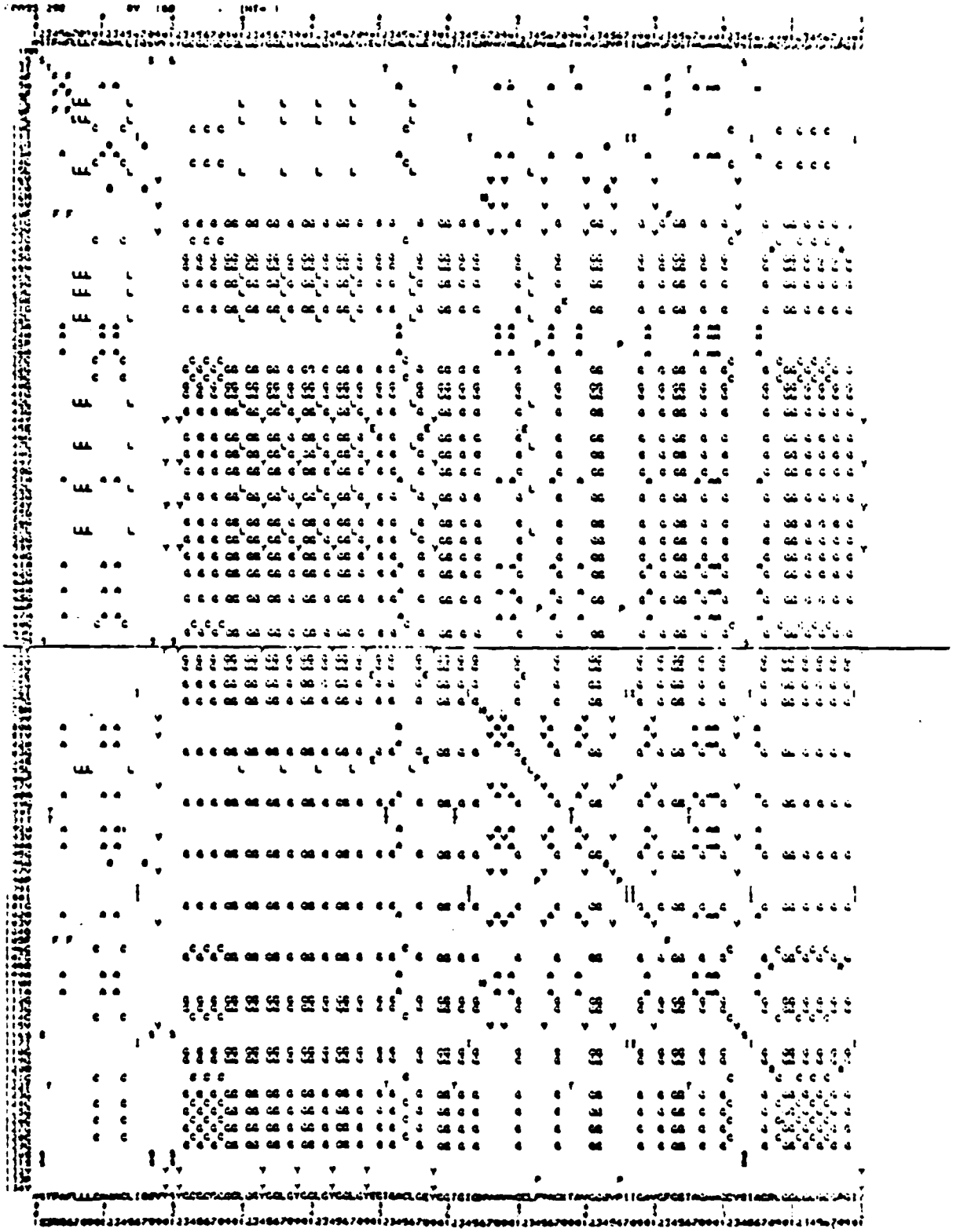


Fig. 2-2. CC matrix for proteins 292 and 187.

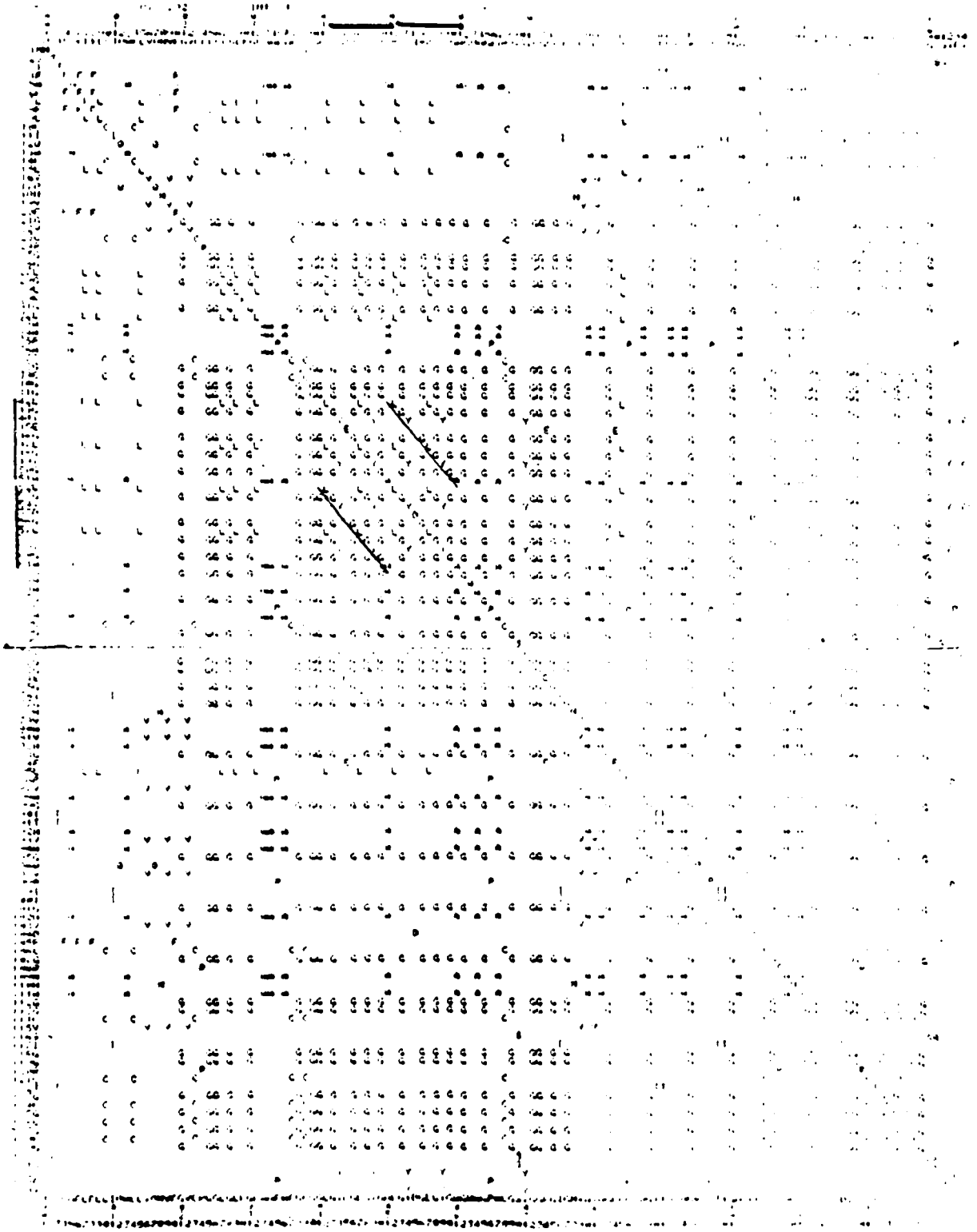


Fig. 2-4. CC matrix for protein 292.

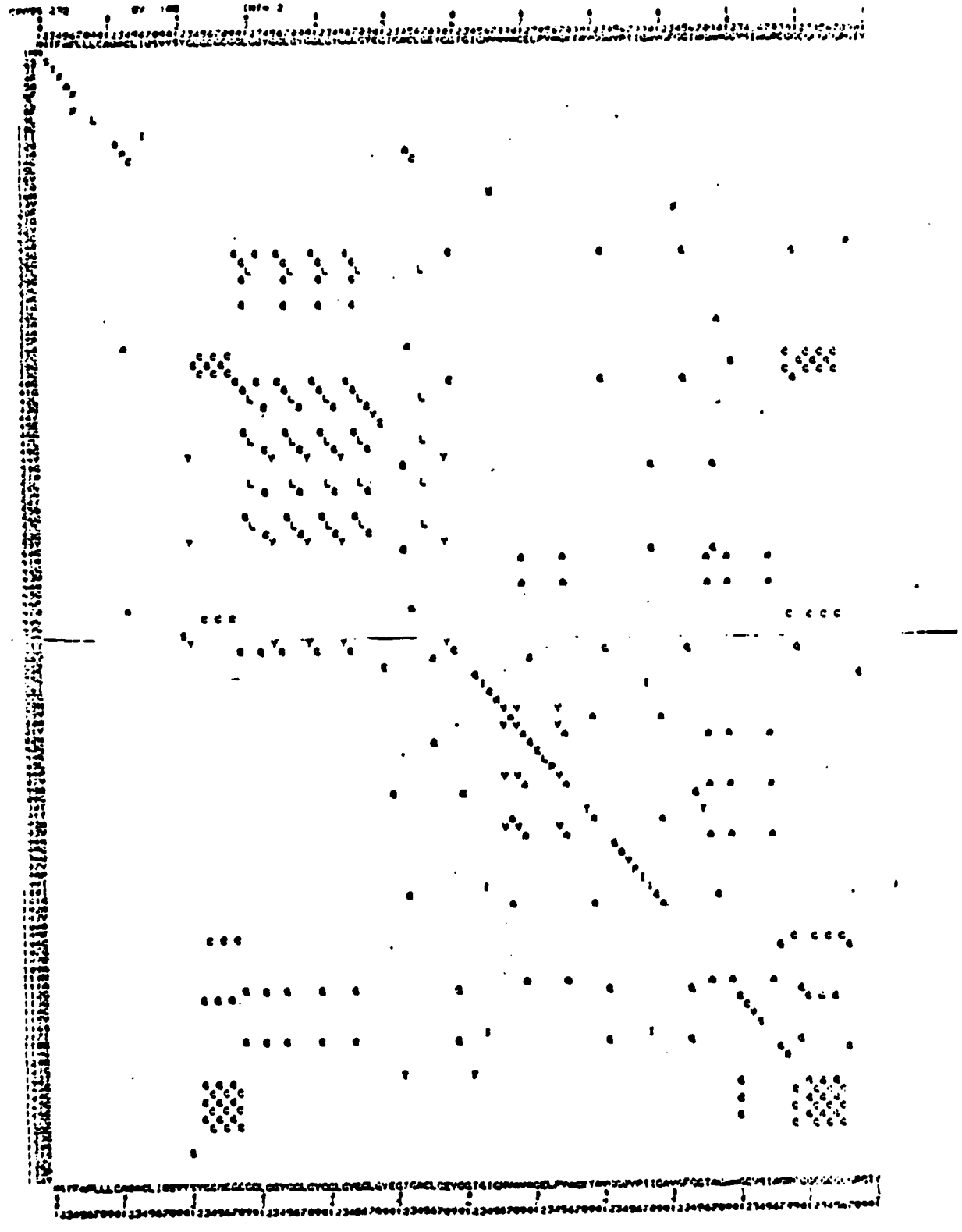


Fig. 2-5. [illegible] matrix for proteins 292 and 188

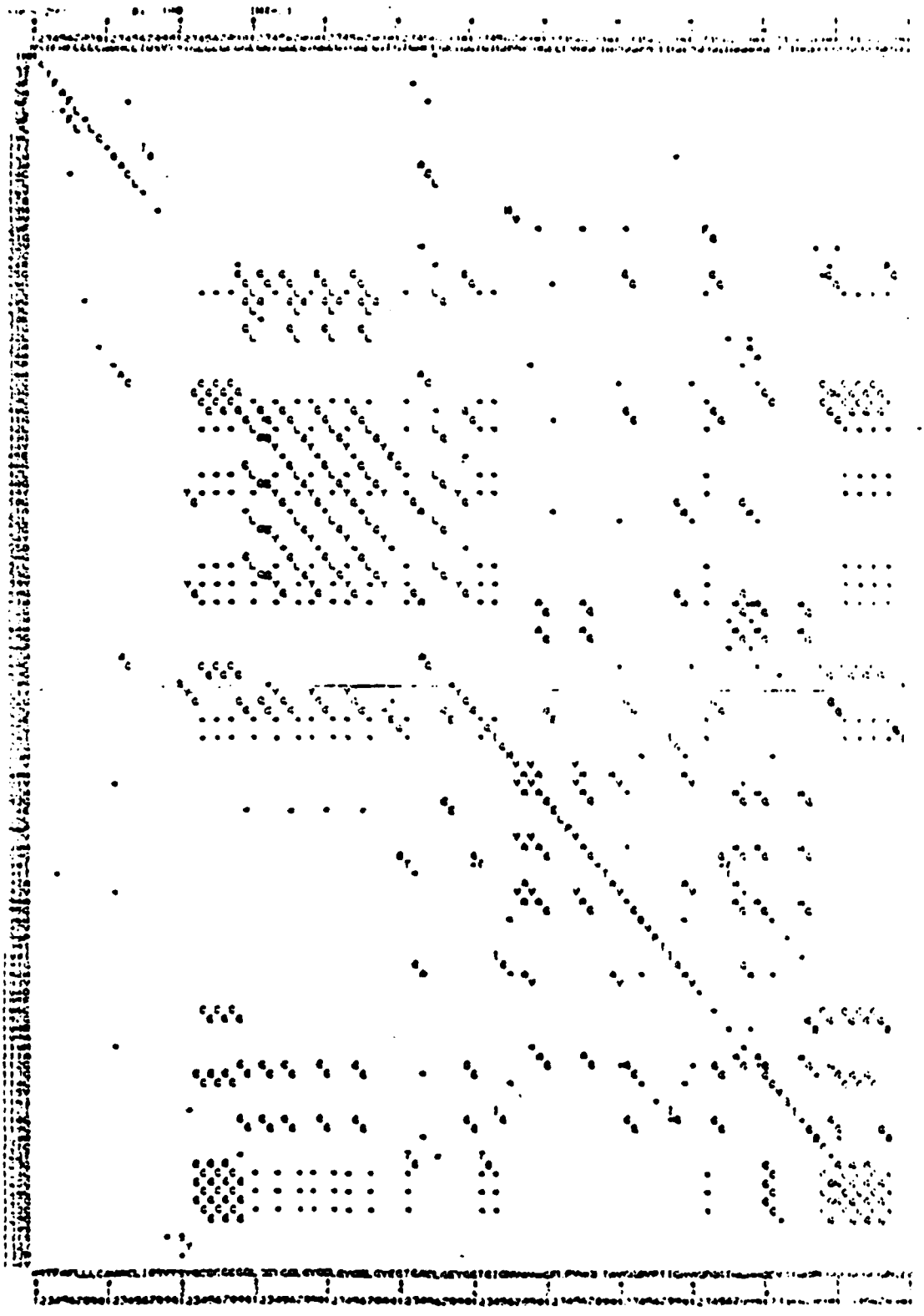


Fig. 2-6. DMC1 matrix for proteins 292 and 183.

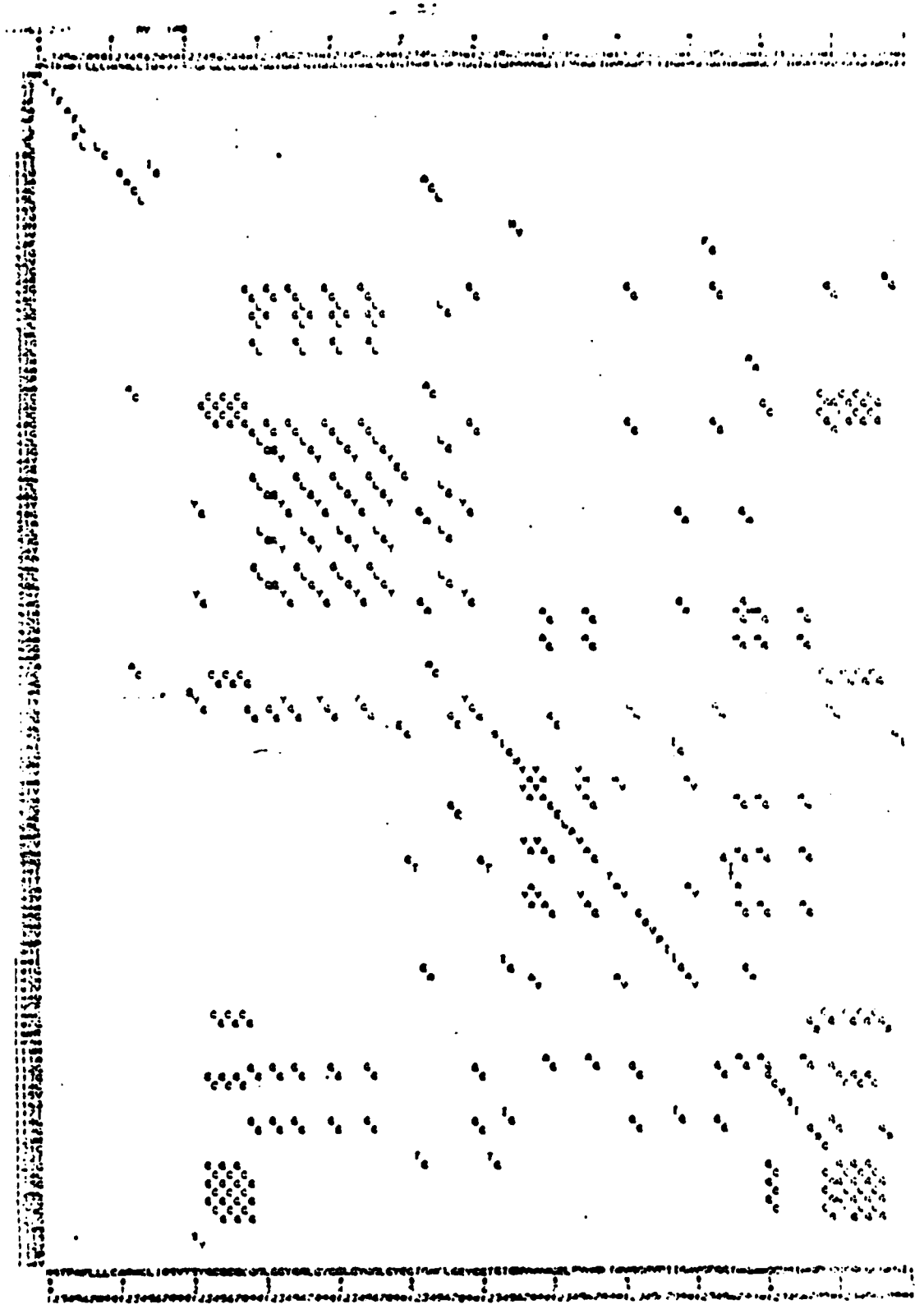


Fig. 2-7. BMC2 matrix for proteins 292 and 183.

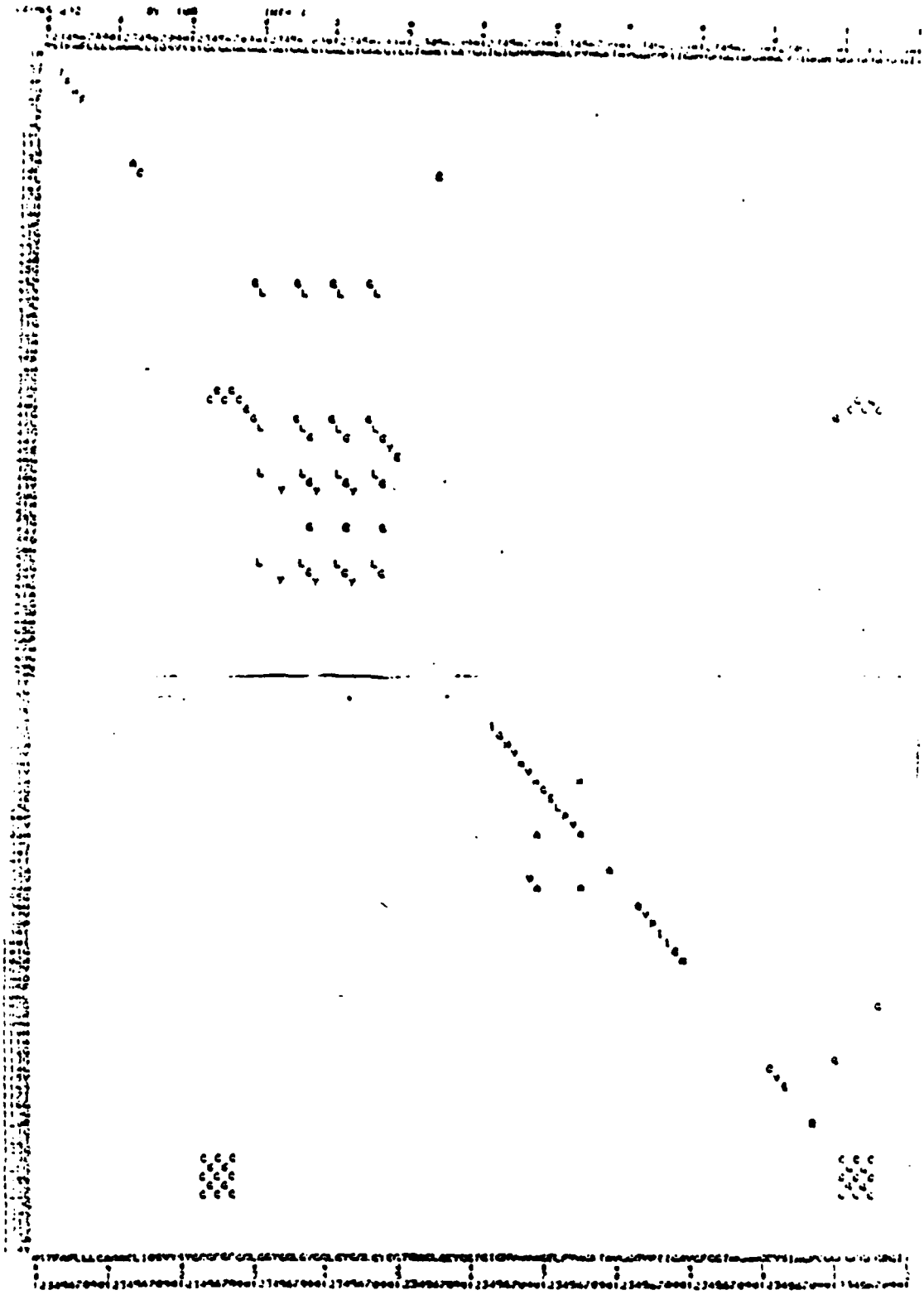


Fig. 2-8. BNC3 matrix for proteins 292 and 138.

3. STATISTICAL PROPERTIES OF SMEARS OF CHARACTER MATRICES

The CC, NNC, PNC, BNC1, BNC2 and BNC3 character matrices were introduced in chapter 2 as graphical tools to explore string data. This chapter derives some of the statistical properties of the smears of the CC and NNC matrices. The statistical properties of matrix smears depend on the model under which words are written. The model the most tractable to work with is the independence model. The independence model supposes that words $\underline{X}=(X_1,X_2,\ldots,X_m)$ and $\underline{Y}=(Y_1,Y_2,\ldots,Y_n)$, written in an alphabet of s characters $\{a_1,\ldots,a_s\}$, are independent sets of independent observations distributed as:

$$\Pr(X_t=a_i)=p_i \qquad i=1,\ldots,s \text{ and } t=1,\ldots,m$$

and $$\Pr(Y_t=a_i)=q_i \qquad i=1,\ldots,s \text{ and } t=1,\ldots,n.$$

Propositions 3-1a and 3-2a derive the first two moments and the asymptotic distributions of the smears of the CC and NNC matrices for two different words $\underline{X}$ and $\underline{Y}$. Propositions 3-1b and 3-2b derive the same results matrices of one word $\underline{X}$.

Proposition 3.1a. Let $S(\underline{X},\underline{Y})$ be the smear of the CC matrix of $\underline{X}$ and $\underline{Y}$ defined in (2-2). Under the independence model,

$$ES(\underline{X},\underline{Y})= \sigma \tag{3-1}$$

where σ is the theoretical smear defined in equation (2-3)

and

$$\mathrm{Var}S(\underline{X},\underline{Y}) = \frac{\sigma}{mn} - \frac{m+n-1}{mn}\sigma^2 + \frac{(n-1)\sum_{k=1}^{s} p_k q_k^2 + (m-1)\sum_{k=1}^{s} p_k^2 q_k}{mn} \tag{3-2}$$

$$- \frac{1}{m}\sum_{k=1}^{s} p_k q_k^2 + \frac{1}{n}\sum_{k=1}^{s} p_k^2 q_k - \left(\frac{1}{m}+\frac{1}{n}\right)\sigma^2 \text{ as } m\to\infty \text{ and } n\to\infty.$$

If $(m/n) \to \lambda$ as $m \to \infty$ and $n \to \infty$,

$$\sqrt{m}(S(\underline{X},\underline{Y}) - \sigma) \xrightarrow{D} N(0,V) \tag{3-3}$$

and

$$V = \sum_{k=1}^{s} p_k q_k^2 + \lambda \sum_{k=1}^{s} p_k^2 q_k - (1+\lambda)\sigma^2 \tag{3-4}$$

Proof. The smear of the CC matrix can be written as:

$$S(\underline{X},\underline{Y}) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i,Y_j)}{mn} \tag{3-5}$$

where

$$\phi(X_i,Y_j) = \begin{cases} 1 & \text{if } X_i = Y_j \\ 0 & \text{otherwise.} \end{cases} \tag{3-6}$$

Therefore,

$$ES(\underline{X},\underline{Y}) = E\phi(X_i,Y_j) = Pr(X_i=Y_j) = \sum_{k=1}^{s} p_k q_k,$$

the RHS of the above equation being the theoretical smear defined in equation (3-3). To compute the variance of $S(\underline{X},\underline{Y})$ we evaluate variances and covariances among the ϕ variables.

$$Var\phi(X_i,Y_j) = \sigma(1-\sigma)$$

If $i \neq u$ and $j \neq v$, $\quad Cov(\phi(X_i,Y_j),\phi(X_u,Y_v)) = 0.$

If $j \neq v$ $\quad Cov(\phi(X_i,Y_j),\phi(X_i,Y_v)) = \sum_{k=1}^{s} p_k q_k^2 - \sigma^2.$

If $i \neq u$ $\quad Cov(\phi(X_i,Y_j),\phi(X_u,Y_j)) = \sum_{k=1}^{s} p_k^2 q_k - \sigma^2.$

Hence

$$m^2n^2\mathrm{Var}S(\underline{X},\underline{Y}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{Var}\phi(X_i,Y_j) + \sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq v}}^{n}\sum_{v=1}^{n}\mathrm{Cov}(\phi(X_i,Y_j),\phi(X_i,Y_v))$$

$$+ \sum_{i=1}^{m}\sum_{\substack{j=1\\ i\neq u}}^{n}\sum_{u=1}^{m}\mathrm{Cov}(\phi(X_i,Y_j),\phi(X_u,Y_j))$$

$$= mn\sigma(1-\sigma)+mn(n-1)\left(\sum_{k=1}^{s} p_k q_k^2 - \sigma^2\right) + mn(m-1)\left(\sum_{k=1}^{s} p_k^2 q_k - \sigma^2\right).$$

Equation (3-2) is obtained by dividing both sides of this equation by $(mn)^2$.

To derive the asymptotic distribution of the smear of the CC matrix, note that equation (3-6) presents $S(\underline{X},\underline{Y})$ as a U-statistic. Therefore if $(m/n) \to \lambda$ as $m \to \infty$ and $n \to \infty$, the asymptotic distribution of S is normal (see, e.g., theorem 9, p.364, in Lehmann [9]. In particular

$$\sqrt{m}(S(\underline{X},\underline{Y}) - \sigma) \xrightarrow{D} N(0,V)$$

for
$$V = \sigma_{01}^2 + \lambda\sigma_{10}^2, \tag{3-7}$$

where $\sigma_{10}^2 = \mathrm{Var}\Phi_{10}(X_t)$, $\sigma_{01}^2 = \mathrm{Var}\Phi_{01}(Y_t)$

and
$$\Phi_{10}(x) = E\phi(x,Y_t) - \sigma = Pr(Y_t = x) - \sigma$$

$$\Phi_{01}(y) = E\phi(X_t,y) - \sigma = Pr(X_t = y) - \sigma.$$

Hence
$$\sigma_{10}^2 = \sum_{k=1}^{s} p_k(q_k-\sigma)^2,$$

$$\sigma_{01}^2 = \sum_{k=1}^{s} q_k(p_k-\sigma)^2,$$

and (3-4) is obtained by substituting the above expressions for σ_{10} and σ_{01} into equation (3-7).

Remark that if $m\to\infty$ and $n\to\infty$ so that $(m/n) \to \lambda$, $m\mathrm{Var}S(\underline{X},\underline{Y}) \to V$,

i.e. the limit of the variance is the variance of the asymptotic distribution of $S(\underline{X},\underline{Y})$.

The asymptotic result of proposition 1 may also be obtained by the δ method. The δ method is used to prove the asymptotic result in proposition 3-1b which can be may also be proved by the 1-sample U-statistics theorem.

Proposition 3-1b. Let $S(\underline{X})$ be the smear of the CC matrix of $\underline{X}$. Under the independence model,

$$ES(\underline{X}) = \frac{1}{m} + (1-\frac{1}{m}) \sum_{k=1}^{s} p_k^2 \sim \sum_{k=1}^{s} p_k^2 \quad \text{as } m \to \infty \tag{3-8}$$

and

$$\mathrm{Var}S(\underline{X}) = 2\frac{m-1}{m^3}\sum_{k=1}^{s} p_k^2 + 4\frac{(m-1)(m-2)}{m^3}\sum_{k=1}^{s} p_k^3 - \frac{2(m-1)(mn-3)}{m^3}(\sum_{k=1}^{s} p_k^2)^2$$

$$- \frac{4}{m}(\sum_{k=1}^{s} p_k^3 - (\sum_{k=1}^{s} p_k^2)^2)^2 \tag{3-9}$$

Furthermore, as $m \to \infty$

$$\sqrt{m}(S(\underline{X}) - \sum_{k=1}^{s} p_k^2) \xrightarrow{\mathcal{D}} N(0,\ 4(\sum_{k=1}^{s} p_k^3 - (\sum_{k=1}^{s} p_k^2)^2)) \tag{3-10}$$

Proof. As noted in chapter 2 a CC matrix for one word carries a solid diagonal throughout. For $\underline{X}=\underline{Y}$, equation 3-5 becomes

$$S(\underline{X}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m}\phi(X_i,X_j)}{m^2} = \frac{1}{m} + \frac{\sum\sum_{i\neq j}\phi(X_i,X_j)}{m^2}. \tag{3-11}$$

Under the independence assumption, for $i\neq j$, $E\phi(X_i,X_j) = \sum_{k=1}^{s} p_k^2$ and equation (3-8) follows by taking expectations of both sides of (3-11).

For notational convenience let

$$\tau = \sum_{k=1}^{s} p_k^2. \tag{3-12}$$

To compute the variance of $S(\underline{X})$ we evaluate covariances among the ϕ's.

$$\text{Cov}(\phi(X_i,X_j),\phi(X_j,X_i) = \text{Var}\phi(X_i,X_j)$$

If $i\neq j$ $\quad \text{Var}\phi(X_i,X_j) = \tau(1-\tau)$.

If $i\neq u, i\neq v, j\neq v$, and $j\neq v$ $\quad \text{Cov}(\phi(X_i,X_j),\phi(X_u,X_v))= 0$.

If $i\neq j, j\neq k, i\neq k$ $\quad \text{Cov}(\phi(X_i,X_j),\phi(X_i,X_k))= E\phi(X_i,X_j)\phi(X_i,X_k)-\tau^2$

$$= \text{Pr}(X_i=X_j=X_k)-\tau^2 = \sum_{k=1}^{s} p_k^3 - \tau^2.$$

Hence,

$$m^4\text{Var}S(\underline{X}) = \sum_{i\neq j}\sum \text{Var}\phi(X_i,Y_j)+\sum_{i\neq j}\sum \text{Cov}(\phi(X_i,X_j),\phi(X_j,X_i))+$$

$$+ \sum_{\substack{i\neq j\\ i\neq k}}\sum_{j\neq k}\sum \text{Cov}(\phi(X_i,X_j),\phi(X_i,X_k) + \sum_{\substack{i\neq j\\ i\neq k}}\sum_{j\neq k}\sum \text{Cov}(\phi(X_i,X_j),\phi(X_k,X_i) +$$

$$+ \sum_{\substack{i\neq j\\ i\neq k}}\sum_{j\neq k}\sum \text{Cov}(\phi(X_i,X_j),\phi(X_k,X_j) + \sum_{\substack{i\neq j\\ i\neq k}}\sum_{j\neq k}\sum \text{Cov}(\phi(X_i,X_j),\phi(X_j,X_k)=$$

$$= 2m(m-1)\tau(1-\tau)+ 4m(m-1)(m-2)(\sum_{k=1}^{s} p_k^3 - \tau^2),$$

and equation (3-9) is obtained by dividing both sides of the above equation by m^4.

To derive the asymptotic distribution of the smear of the CC matrix by the δ method, let M_i be the count of alphabet character a_i in $\underline{X}=(X_1,\ldots,X_m)$, and let $\hat{p}_i=(M_i/m)$ be the frequency of characte- a_i in $\underline{X}$ and $\hat{p}^T = (\hat{p}_1,\ldots,\hat{p}_s)$. Then $(M_1,\ldots,M_s)$ is multinomially distributed with parameters m and $p^T=(p_1,\ldots p_s)$. By the normal approximation to the multinomial distribution, $\sqrt{m}(\hat{p}-p) \rightarrow N(0,D_p-pp^T)$, where D_p is the

diagonal matrix with the entries of vector $\underline{p}$ along the diagonal, and consequently $\|\hat{\underline{p}}-\underline{p}\|=O_p(1/\sqrt{m})$.

Let $g(\underline{p})=\sum_{k=1}^{s} p_k^2$. From equation (2-4) it follows that $S(\underline{X}) = g(\hat{\underline{p}})$. Then

$$g(\hat{\underline{p}})=g(\underline{p})+(\text{grad } g)^T.(\hat{\underline{p}}-\underline{p})+\varepsilon_m\|\hat{\underline{p}}-\underline{p}\|$$

with $\varepsilon_m\to 0$ as $\hat{\underline{p}}\to\underline{p}$.

Substituting grad $g= 2\underline{p}$ and multiplying the above expansion by $\sqrt{m}$ we obtain

$$\sqrt{m}(g(\hat{\underline{p}})-g(\underline{p})) = \sqrt{m}\ 2\underline{p}^T(\hat{\underline{p}}-\underline{p})+ o_p(1).$$

Therefore, $\sqrt{m}(g(\hat{\underline{p}})-g(\underline{p}))$ is asymptotically normally distributed with mean 0 and variance $4\underline{p}^T(D_p-\underline{p}\underline{p}^T)\underline{p} = 4(\underline{p}^TD_p\underline{p}-(\underline{p}^T\underline{p})^2)$ which is written in the entries of $\underline{p}$ in equation (3-10).

Proposition 3-2a. Let $S(\underline{X},\underline{Y})$ be the smear of the NNC matrix for words $\underline{X}$ and $\underline{Y}$. Under the independence model,

$$ES(\underline{X},\underline{Y})= \sigma^2, \tag{3-13}$$

and as $m\to\infty$ and $n\to\infty$ subject to $m=o(n^2)$ and $n=o(m^2)$,

$$\begin{aligned}\text{Var}S(\underline{X},\underline{Y})\sim\ & \frac{1}{m}\{2\sigma^2(\sum_{k=1}^{s} p_kq_k^2) + (\sum_{k=1}^{s} p_kq_k^2)^2\} \\ & + \frac{1}{n}\{2\sigma^2(\sum_{k=1}^{s} p_k^2q_k) + (\sum_{k=1}^{s} p_k^2q_k)^2\} - 3\ (\frac{1}{m}+\frac{1}{n})\ \sigma^4.\end{aligned} \tag{3-14}$$

If $(m/n)\to\lambda$ as $m\to\infty$ and $n\to\infty$, the smear is asymptotically normally distributed

$$\sqrt{m}(S(\underline{X},\underline{Y})- \sigma^2) \xrightarrow{D} N(0,V)$$

with $$V = \lim m\text{Var}S(\underline{X},\underline{Y}) \text{ as } m\to\infty.$$

Proof. The smear of the NNC matrix for $\underline{X}$ and $\underline{Y}$ can be written as

$$S(\underline{X},\underline{Y})=\frac{\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\Phi(X_i,X_{i+1};Y_j,Y_{j+1})}{(m-1)(n-1)} \tag{3-15}$$

where

$$\Phi(X_i,X_{i+1};Y_j,Y_{j+1})=\begin{cases}1 & \text{if } X_i=Y_j \text{ and } X_{i+1}=Y_{j+1}\\ 0 & \text{otherwise.}\end{cases} \tag{3-16}$$

Therefore,

$$ES(\underline{X},\underline{Y})=E\Phi(X_i,X_{i+1};Y_j,Y_{j+1})=Pr(X_i=Y_j,X_{i+1}=Y_{j+1})=(Pr(X_i=Y_j))^2=\sigma^2,$$

and

$$\text{Var}S(\underline{X},\underline{Y})$$
$$=\frac{1}{(m-1)^2(n-1)^2}\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}\sum_{u=1}^{m-1}\sum_{v=1}^{n-1}\text{Cov}(\Phi(X_i,X_{i+1};Y_j,Y_{j+1}),\Phi(X_u,X_{u+1};Y_v,Y_{v+1})). \tag{3-17}$$

To evaluate (3-17) let

$$V_{ij}=\sum_{k=1}^{m-1}\sum_{l=1}^{n-1}\text{Cov}(\Phi(X_i,X_{i+1};Y_j,Y_{j+1}),\Phi(X_u,X_{u+1};Y_v,Y_{v+1})) \tag{3-18}$$

and rewrite equation (3-17) as

$$\text{Var}S(\underline{X},\underline{Y})=\frac{1}{(m-1)^2(n-1)^2}\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}V_{ij}. \tag{3-19}$$

We evaluate V_{ij} for i=2,....,m-1 and j=2,....,n-1.

By independence, if $|i-u|>1$ and $|j-v|>1$

$$\text{Co}(\Phi(X_i,X_{i+1};Y_j,Y_{j+1}),\Phi(X_u,X_{u+1};Y_v,Y_{v+1}))=0.$$

Therefore the values of u and v which contribute to V_{ij} are such that either $|i-u|\leq 1$ or $|j-v|\leq 1$ as presented in figure 3-1. We now compute their contributions.

If u=i and $|v-j|>1$,

$$\text{Co}(\Phi(X_i,X_{i+1};Y_j,Y_{j+1}),\Phi(X_u,X_{u+1};Y_v,Y_{v+1}))=$$
$$=E\Phi(X_i,X_{i+1};Y_j,Y_{j+1})\Phi(X_i,X_{i+1};Y_v,Y_{v+1})-\sigma^4$$

$$= \Pr(X_i=Y_j=Y_v, X_{i+1}=Y_{j+1}=Y_{v+1}) - \sigma^4 =$$

$$= \left(\sum_{k=1}^{s} p_k q_k^2\right)^2 - \sigma^4. \qquad (3\text{-}20)$$

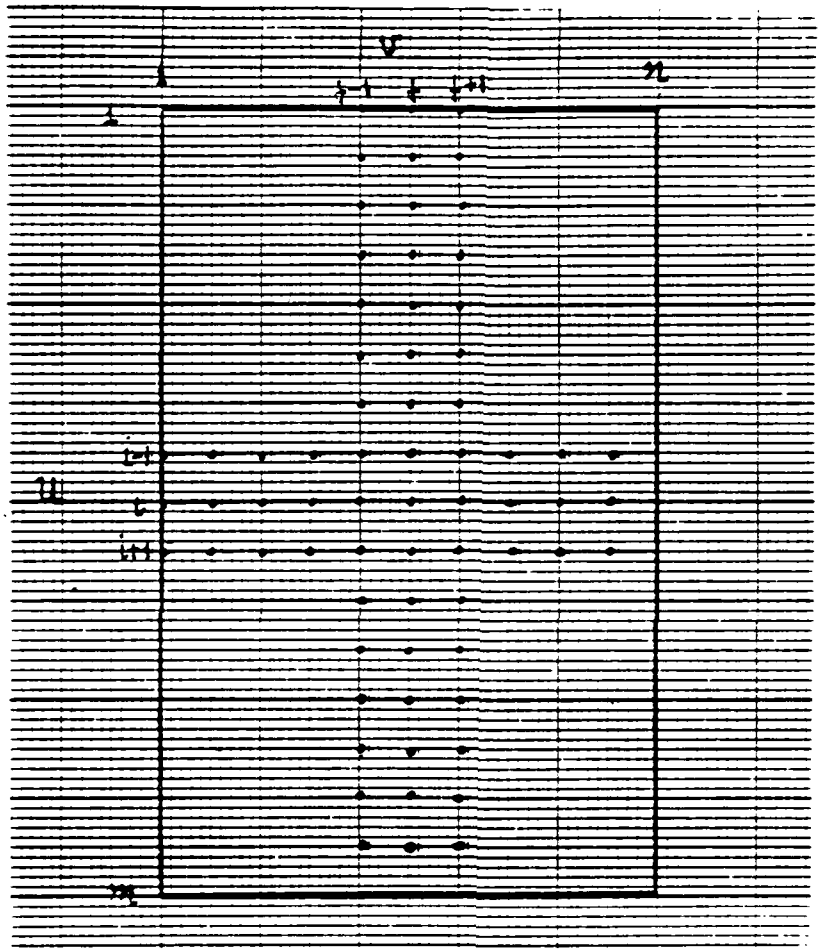


Fig. 3-1. Values of u and v for which

$\mathrm{Cov}(\Phi(X_i, X_{i+1}; Y_j, Y_{j+1}), \Phi(X_u, X_{u+1}; Y_v, Y_{v+1})) = 0$ for i and j fixed.

If $u = i \pm 1$ and $|j-v| > 1$,

$$\mathrm{Cov}(\Phi(X_i, X_{i+1}; Y_j, Y_{j+1}), \Phi(X_u, X_{u+1}; Y_v, Y_{v+1})) = \sigma^2\left(\sum_{k=1}^{s} p_k q_k^2 - \sigma^2\right). \qquad (3\text{-}21)$$

Similarly, if $v = j$ and $|i-u| > 1$

$$\mathrm{Cov}(\Phi(X_i,X_{i+1};Y_j,Y_{j+1}),\Phi(X_u,X_{u+1};Y_v,Y_{v+1})) = (\sum_{k=1}^{\bar{s}} p_k^2 q_k)^2-\sigma^4, \qquad (3\text{-}22)$$

and if $v=j\pm1$ and $|i-u|>1$,

$$\mathrm{Cov}(\Phi(X_i,X_{i+1};Y_j,Y_{j+1}),\Phi(X_u,X_{u+1};Y_v,Y_{v+1})) = \sigma^2(\sum_{k=1}^{\bar{s}} p_k^2 q_k-\sigma^2). \qquad (3\text{-}23)$$

As indicated in figure 3-1 for fixed $i=2,\ldots,m-1$ and $j=2,\ldots,n-1$, V_{ij} is a sum of $3(m-4)+3(n-4)-9$ nonzero covariances. Of those all but nine can be computed from formulae (3-20) to (3-23), the nine terms being

$$\mathrm{Cov}(\Phi(X_i,X_{i+1};Y_j,Y_{j+1}),\Phi(X_u,X_{u+1};Y_v,Y_{v+1}))$$

for u and v such that $|u-i|\leq1$ or $|j-v|\leq1$. These covariances can be derived similarly. However, we do not need to compute them explicitly as their total contribution to V_{ij} is O(1) compared with that of the other contributing terms which may be computed from formulae (3-20) to (3-23) and is of order $O(m)+O(n)$ as can be seen from (3-24) below. Thus,

$$\begin{aligned} V_{ij} = & \sum_{|v-j|>1}^{\bar{}} \mathrm{Cov}(\Phi(X_i,X_{i+1};Y_j,Y_{j+1},\Phi(X_i,X_{i+1};Y_v,Y_{v+1}))+ \\ & + \sum_{u=i\pm1}^{\bar{}} \sum_{|v-j|>1}^{\bar{}} \mathrm{Cov}(\Phi(X_i,X_{i+1};Y_j,Y_{j+1}),\Phi(X_u,X_{u+1};Y_v,Y_{v+1})) + \\ & + \sum_{|u-i|>1}^{\bar{}} \mathrm{Cov}(\Phi(X_i,X_{i+1};Y_j,Y_{j+1}),\Phi(X_u,X_{u+1};Y_j,Y_{j+1}))+ \\ & + \sum_{|u-i|>1}^{\bar{}} \sum_{v=j\pm1}^{\bar{}} \mathrm{Cov}(\Phi(X_i,X_{i+1};Y_j,Y_{j+1}),\Phi(X_u,X_{u+1};Y_v,Y_{v+1})+ O(1). \end{aligned}$$

Each of the four sums in the above equation can be computed by substituting formulae in (3-21) to (3-24) for the covariances in the four sums above to obtain

$$\begin{aligned} V_{ij}= & \ 2(n-4)\sigma^2(\sum_{k=1}^{\bar{s}} p_k q_k^2-\sigma^2) + (n-4)((\sum_{k=1}^{\bar{s}} p_k q_k^2)^2-\sigma^4) \\ & +2(m-4)\sigma^2(\sum_{k=1}^{\bar{s}} p_k^2 q_k-\sigma^2) + (m-4)((\sum_{k=1}^{\bar{s}} p_k^2 q_k)^2-\sigma^4) + O(1). \qquad (3\text{-}24) \end{aligned}$$

The value of V_{ij} when i=1 or j=1 is slightly different from the expression in 3-2. However,

$\sum_{i=2}^{m-1}\sum_{j=2}^{n-1} V_{ij}$ is of order $O(m^2n)+O(mn^2)$ whereas $\sum_{j=1}^{n-1} V_{1j}$ and $\sum_{i=1}^{m-1} V_{i1}$

Hence, as $m\to\infty$ and $n\to\infty$ subject to $m=o(n^2)$ and $n=o(m^2)$,

$$\begin{aligned}\mathrm{Var}S(\underline{X},\underline{Y}) &\sim \frac{1}{(m-1)^2}\frac{1}{(n-1)^2}\sum_{i=2}^{m-1}\sum_{j=2}^{n-1} V_{ij} \\ &\sim \frac{1}{n}\left(\sigma^2\left(2\sum p_k^2 q_k-\sigma^2\right)+\left(\sum p_k^2 q_k\right)^2-\sigma^4\right) \\ &+ \frac{1}{n}\left(\sigma^2\left(2\sum p_k q_k^2-\sigma^2\right)+\left(\sum p_k q_k^2\right)^2-\sigma^4\right) \qquad (3\text{-}25)\end{aligned}$$

which equals the MIS of equation (3-14).

To derive the asymptotic distribution of $S(\underline{X},\underline{Y})$ let M_{ij} and N_{ij} be the counts of the 2-letter syllable a_ia_j in $\underline{X}$ and $\underline{Y}$ respectively. Formally,

$$M_{ij}=\sum_{k=1}^{m-1} I_{ij}(X_k,X_{k+1}) \qquad N_{ij}=\sum_{k=1}^{n-1} I_{ij}(Y_k,Y_{k+1}) \qquad (3\text{-}26)$$

for $I_{ij}(...)$ the indicator function

$$I_{ij}(x,y)=\begin{cases}1 & \text{if } x=a_i \text{ and } y=a_j\\ 0 & \text{otherwise.}\end{cases} \qquad (3\text{-}27)$$

Note that

$$\sum_{i=1}^{s}\sum_{j=1}^{s} M_{ij} = m-1 \qquad \sum_{i=1}^{s}\sum_{j=1}^{s} N_{ij} = n-1$$

and let

$$\hat{p}_{ij}=\frac{M_{ij}}{m-1} \quad \text{and} \quad \hat{q}_{ij}=\frac{N_{ij}}{n-1} \qquad (3\text{-}28)$$

be the frequencies of syllable a_ia_j in $\underline{X}$ and $\underline{Y}$. The following notation

is useful:

$$M^T=(M_{11},\ldots,M_{1s},\ldots,M_{s1},\ldots,M_{ss}) \tag{3-29}$$

$$N^T=(N_{11},\ldots,N_{1s},\ldots,N_{s1},\ldots,N_{ss})$$

$$p^T=(p_{11},\ldots,p_{1s},\ldots,p_{s1},\ldots,p_{ss}) \tag{3-30}$$

$$q^T=(q_{11},\ldots,q_{1s},\ldots,q_{s1},\ldots,q_{ss})$$

$$\hat{p}^T=(\hat{p}_{11},\ldots,\hat{p}_{1s},\ldots,\hat{p}_{s1},\ldots,\hat{p}_{ss}) \tag{3-31}$$

$$\hat{q}^T=(\hat{q}_{11},\ldots,\hat{q}_{1s}\ \ldots,\hat{q}_{s1},\ldots,\hat{q}_{ss}).$$

In this notation, the smear of the NNC character matrix can be written as

$$S(\underline{X},\underline{Y})=\frac{\sum_{i=1}^{s}\sum_{j=1}^{s}M_{ij}N_{ij}}{(m-1)(n-1)} = \hat{p}^T.\hat{q}. \tag{3-32}$$

Let $$g(p,q) = p^T.q.$$

By the differentiability of $g(\ ,\ )$,

$$g(\hat{p},\hat{q}) = g(p,q) +(\text{grad } g)^T.\ \left(\begin{bmatrix}\hat{p}\\ \hat{q}\end{bmatrix}-\begin{bmatrix}p\\ q\end{bmatrix}\right) + \varepsilon_{mn}.\left\|\begin{bmatrix}\hat{p}-p\\ \hat{q}-q\end{bmatrix}\right\|$$

where $\varepsilon_{mn}\to 0$ as $\hat{p}\to p$ and $\hat{q}\to q$. Substituting

$$(\text{grad } g)^T= \left(\frac{\partial}{\partial p_{11}},\ldots,\frac{\partial}{\partial p_{ss}},\ldots,\frac{\partial}{\partial q_{11}},\ldots,\frac{\partial}{\partial q_{ss}}\right)\ g(p,q)= \begin{bmatrix}q\\ p\end{bmatrix}$$

into the above equation, we obtain

$$S(\underline{X},\underline{Y})= g(p,q) + q^T.(\hat{p}-p) + p^T.(\hat{q}-q) + \varepsilon_{mn}.\left\|\begin{bmatrix}\hat{p}-p\\ \hat{q}-q\end{bmatrix}\right\|. \tag{3-33}$$

Note that the distribution of M and N is not multinomial. The asymptotic distibution of $\hat{p}$ (and $\hat{q}$) is provided by the following lemma.

Lemma 3.1. Under the independence model, $\hat{p}$ defined by (3-31) and (3-28) is asymptotically normally distributed

$$\sqrt{m}\ (\hat{p} - E\hat{p}) \xrightarrow{D} N(0,\Sigma^1)$$

with

$$E\underline{p}=(p_1^2,\ldots,p_1p_s,\ldots,p_sp_1,\ldots,p_s^2) \tag{3-34}$$

$$\Sigma^1_{i,j;u,v}= p_ip_jp_u\delta_{iv} + p_ip_j\delta_{iu}\delta_{jv} + p_ip_jp_v\delta_{ju} - 3p_ip_jp_up_v \tag{3-35}$$

(δ_{ij} is the usual Kronecker delta.)

Proof. Let $c^T=(c_{11},\ldots,c_{1s},\ldots,c_{s1},\ldots,c_{ss})$ be an arbitrary vector of s^2 constants. By (3-29) and (3-26),

$$c^TM = \sum_{i=1}^{s}\sum_{j=1}^{s} c_{ij}M_{ij} = \sum_{\alpha=1}^{m-1}\sum_{i=1}^{s}\sum_{j=1}^{s} c_{ij}I_{ij}(X_\alpha,X_{\alpha+1})$$

Note that c^TM is an $(m-1)$st sum of 1-dependent variables.

$$\mathrm{Var}(c^TM) = \sum_i\sum_j\sum_u\sum_v c_{ij}c_{uv}\mathrm{Cov}(M_{ij},M_{uv})$$

and $$\mathrm{Cov}(M_{ij},M_{uv})= \mathrm{Cov}\Big(\sum_{\alpha=1}^{m-1} I_{ij}(X_\alpha,X_{\alpha+1}),\sum_{\beta=1}^{m-1} I_{uv}(X_\beta,X_{\beta+1})\Big).$$

If $|\alpha-\beta|>1$ by independence, $\mathrm{Cov}(I_{ij}(X_\alpha,X_{\alpha+1}),I_{uv}(X_\beta,X_{\beta+1}))=0$.

Hence,

$$\begin{aligned}\mathrm{Cov}(M_{ij},M_{uv})= &\sum_{\alpha=2}^{m-1}\mathrm{Cov}(I_{ij}(X_\alpha,X_{\alpha+1}),I_{uv}(X_{\alpha-1},X_\alpha))\\ &+\sum_{\alpha=1}^{m-1}\mathrm{Cov}(I_{ij}(X_\alpha,X_{\alpha+1}),I_{uv}(X_\alpha,X_{\alpha+1}))\\ &+\sum_{\alpha=1}^{m-2}\mathrm{Cov}(I_{ij}(X_\alpha,X_{\alpha+1}),I_{uv}(X_{\alpha+1},X_{\alpha+2})).\end{aligned} \tag{3-36}$$

The covariances in the three sums above are:

$$\begin{aligned}&\mathrm{Cov}((I_{ij}(X_\alpha,X_{\alpha+1}),I_{uv}(X_{\alpha-1},X_\alpha))=\\ &\quad= \Pr(X_{\alpha-1}=a_u,X_\alpha=a_v,X_\alpha=a_i,X_{\alpha+1}=a_v)-p_ip_jp_up_v\\ &\quad= p_ip_jp_u\delta_{iv}-p_ip_jp_up_v\end{aligned}$$

and similarly,

$$\mathrm{Cov}(I_{ij}(X_\alpha,X_{\alpha+1}),I_{uv}(X_\alpha,X_{\alpha+1})) = p_ip_j\delta_{iu}\delta_{jv} - p_ip_jp_up_v$$

and $$\mathrm{Cov}(I_{ij}(X_\alpha,X_{\alpha+1}),I_{uv}(X_{\alpha+1},X_{\alpha+2})) = p_ip_jp_v\delta_{ju} - p_ip_jp_up_v.$$

Substituting the above formulae for the covariances into (3-35) we obtain

$$\mathrm{Cov}(M_{ij},M_{uv}) = (m-2)(p_ip_jp_u\delta_{iv}+p_ip_jp_v\delta_{ju}-p_ip_jp_up_v) + (m-1)(p_ip_j\delta_{iu}\delta_{jv}-p_ip_jp_up_v). \qquad (3\text{-}37)$$

Therefore $\mathrm{Cov}(M_{ij},M_{uv})=O(m)$ and so is $\mathrm{Var}(c^TM)$. According to the k-dependent CLT (theorem 7.3.1 of Chung [3]),

$$\frac{c^TM - c^TEM}{\sqrt{\mathrm{Var}c^TM}} \to N(0,1)$$

and consequently $$\sqrt{m}(\hat{p} - E\hat{p}) \sim N(0,\Sigma^1),$$

the parameters of the distribution given by (3-34) and (3-35).

Lemma 1 makes the δ method applicable to the expansion (3-33) and the asymprotic distribution of the smear of the MNC matrix easily derivable. If $(m/n)\to\lambda$ as $m\to\infty$ and $n\to\infty$, for $p = E\hat{p}$ and $q = E\hat{q}$ equation (3-33) becomes:

$$\sqrt{m}\,(S(\underline{X},\underline{Y})-\sigma^2) = \sqrt{m}\,E\hat{q}^T.(\hat{p}-E\hat{p}) + \lambda\sqrt{n}\,E\hat{p}^T.(\hat{q}-q) + o_p(1).$$

Therefore the LHS is asymptotically normally distributed with mean 0 and variance

$$E\hat{q}^T\Sigma^1E\hat{q} + \lambda E\hat{p}^T\Sigma^2E\hat{p}.$$

$E\hat{p}$ and Σ^1 are given in (3-34) and (3-35) of lemma 3-1 and the formulae for $E\hat{q}$ and Σ^2 are obtained by interchanging q_k for p_k. Substituting $E\hat{p}$, $E\hat{q}$, Σ^1 and Σ^2 into the above asymptotic variance we obtain

$$2\sigma^2\sum_{k=1}^{s}p_kq_k^2 + (\sum_{k=1}^{s}p_kq_k^2)^2 + \lambda\{2\sigma^2\sum_{k=1}^{s}p_k^2q_k + (\sum_{k=1}^{s}p_k^2q_k)^2\} - 3(1+\lambda)\sigma^4.$$

as asserted in proposition 3-2a.

Remark, again, that the limit of the variance of $\sqrt{m}(S(\underline{X},\underline{Y})-\sigma^2)$ equals the variance of the asymptotic distribution.

Proposition 3-2b. Let $S(\underline{X})$ be the smear of the NNC matrix for $\underline{X}$ written under independence. Then,

$$ES(\underline{X}) = \frac{1}{m-1} + \frac{2(m-2)}{(m-1)^2} \sum_{k=1}^{s} p_k^3 + \frac{(m-2)(m-3)}{(m-1)^2} \tau^2 \sim \tau^2 \qquad (3\text{-}38)$$

and

$$\sqrt{m}(S(\underline{X})-\tau^2) \xrightarrow{D} N(0,\ 4(\ 2\tau^2 \sum_{k=1}^{s} p_k^3 + (\sum_{k=1}^{s} p_k^3)^2 - 3\tau^4)\ ) \qquad (3\text{-}39)$$

Proof. The smear of the NNC matrix for $\underline{X}$ can be written as

$$S(\underline{X}) = \frac{\sum_{i=1}^{m-1}\sum_{j=1}^{m-1} \Phi(X_i,X_{i+1};X_j,X_{j+1})}{(m-1)^2} = \frac{1}{m-1} + \frac{\sum\sum_{i\neq j} \Phi(X_i,X_{i+1};X_j,X_{j+1})}{(m-1)^2} \qquad (3\text{-}40)$$

for $\Phi(.,.,;.,.,)$ defined in (3-16). To evaluate $ES(\underline{X})$ remark that

$$\text{if } |i-j|>1 \quad E\Phi(X_i,X_{i+1};X_j,X_{j+1}) = P(X_i=X_j)P(X_{i+1}=X_{j+1}) = \tau^2 \qquad (3\text{-}41)$$

$$\text{if } |i-j|=1 \quad E\Phi(X_i,X_{i+1};X_j,X_{j+1}) = P(X_i=X_{i+1}=X_{i+2}) = \sum_{k=1}^{s} p_k^3. \qquad (3\text{-}42)$$

$E\Phi$ is given by (3-42) for $2(m-2)$ of the $(m-1)^2-(m-1)=(m-1)(m-2)$ pairs in the summation of (3-40) and by (3-41) for the remaining pairs. Equation (3-38) is obtained by taking expectations on both sides of (3-40) and substituting from (3-41) and (3-42) into (3-40).

To derive the asymptotic distribution of $S(\underline{X})$ note that

$$S(\underline{X}) = \frac{\sum_{i=1}^{s}\sum_{j=1}^{s} M_{ij}^2}{(m-1)^2} = \|\hat{\underline{p}}\|^2$$

for M_{ij} defined by (3-26) and $\hat{\underline{p}}$ defined by (3-31) and (3-28).

Let $g(p)=\|p\|^2$. Since

$$(\text{grad } g)^T=(\frac{\partial}{\partial p_{11}},\ldots,\frac{\partial}{\partial p_{ss}})^T g(p)=2p^T,$$

$$S(\underline{X})=g(p) + 2p^T.(\hat{\underline{p}}-p)+\varepsilon_m\|\hat{\underline{p}}-p\|.$$

For $p=E\hat{\underline{p}}$

$$g(p)=\sum_{i=1}^{s}\sum_{j=1}^{s}(p_ip_j)^2 = \tau^2$$

and $$\sqrt{m}(S(\underline{X})-\tau^2) =2\sqrt{m}\ p^T.(\hat{\underline{p}}-p) + o_p(1).$$

By lemma 3.1 the LHS of the above equation is asymptotically normal with mean 0 and variance which, computed from (3-35), is

$$4\sum_i\sum_j\sum_u\sum_v p_ip_jp_up_v(p_ip_jp_u\delta_{iv}+p_ip_j\delta_{iu}\delta_{jv}+p_ip_jp_v\delta_{ju}-3p_ip_jp_up_v)=$$

$$=4(\sum_i\sum_j\sum_u p_i^3p_j^2p_u^2 + \sum_i\sum_j p_i^3p_j^3 + \sum_i\sum_j\sum_v p_i^2p_j^3p_v^2 -3\tau^4)=$$

$$4(2\sum_{k=1}^{s} p_k^3\tau^2 + (\sum_{k=1}^{s} p_k^3)^2-3\tau^4)$$

as asserted in (3-39).

If $\underline{X}$ and $\underline{Y}$ are two independent identically distributed words, i.e., if $m=n$ and $p_i=q_i$ for all i, both $S(\underline{X})$ and $S(\underline{X},\underline{Y})$ estimate the same parameter τ of equation (3-12). Propositions 3-1 and 3-2 assert that the variance of the asymptotic distribution of $\sqrt{m}(S(\underline{X})-\tau)$ is twice that of $\sqrt{m}(S(\underline{X},\underline{Y})-\tau)$ for both the crude and the NNC character matrices.

4. SMEARS ALONG DIAGONALS OF CHARACTER MATRICES FOR STRING DATA

Chapter 2 introduced a variety of character matrices that are very useful in bringing out visually similarities among different words or within a word. The visual examination of character matrices - as insightful as it can be - is only a first step in the analysis of string data as it is limited in two aspects.

(i) It is stressful to the investigator's eye.

(ii) While bringing out strings that may be shared between the words under comparison, it falls short of assessing similarities quantitatively. As a consequence, visual recognition of common strings is partially subjective.

This chapter addresses the question of how to make the detection of diagonals objective, i.e. describable quantitatively, and possible to implement on a machine. We are looking for statistics which reflect the presence of diagonals in character matrices.

The statistic that has attracted the attention of researchers so far is the length of the longest common subsequence (LLCS) of the two words under examination. For $\underline{X}=(X_1,\ldots,X_m)$ and $\underline{Y}=(Y_1,\ldots,Y_n)$, the LLCS can be defined as:

$\max\{k: \; X_{i_1}=Y_{j_1},\ldots,X_{i_k}=Y_{j_k} \text{ for } 1\leq i_1<\ldots<i_k\leq m \text{ and } 1\leq j_1<\ldots<j_k\leq n\}$.

Needleman and Wunch [11] were the first to propose the LLCS as a measure of similarity between genetic sequences. Their method to find the LLCS was later modified to a more efficient dynamic programming algorithm by Sankoff [13]. After the LLCS has been computed, the path $\{(i_h,j_h): 1\leq h\leq k\}$ through the crude character matrix of $\underline{X}$ and $\underline{Y}$ can be traced for the

investigator to examine.

The algorithm of Needleman and Wunch suffers from three drawbacks. If a relatively long string is present in both $\underline{X}$ and $\underline{Y}$, it will most probably contribute to the LLCS. However, if the common string is present once in $\underline{X}$ and in two repeats in $\underline{Y}$ as shown in figure 4-1, then of the two different common subsequences of approximately equal length, the algorithm will only select one and will not let the molecular biologist know of the internal repeat.

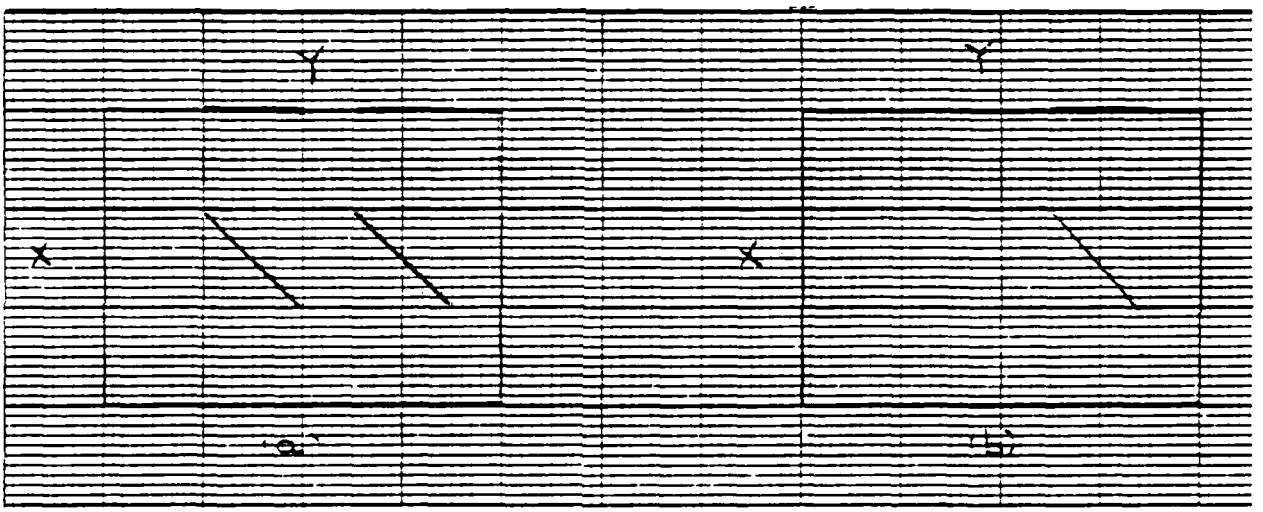


Fig. 4-1. The presence of a relatively long string in both $\underline{X}$ and $\underline{Y}$ in (a) will be most probably detected by the Needleman-Wunch algorithm. A repeat of the string within $\underline{Y}$ as in (b) will not be detected.

Furthermore, the algorithm weighs all matches and mismatches in the same way, counter to the sense of the statistician that matches in rare letters should weigh more than matches among rather frequent letters, and the knowledge of the molecular biologist that some substitutions on genetic molecules affect the function of the molecules and the state of their cells dramatically, whereas others do not. Needleman and Wunch

were aware of this problem and mentioned that weights other than 0 and 1 could be used but they provided no hints as to how to obtain these weights from data and their remark was later ignored in the mathematical literature. Finally, little is known of the distributional properties of the LLCS. Chvatal and Sankoff [4] showed that if L_n is the LLCS between two words both of lengh n, EL_n is superadditive with respect to n, i.e. $EL_{m+n} \geq EL_m+EL_n$, and therefore $E(L_n/n)$ converges to a constant that depends on the size of the alphabet in which the words are written. They also provided upper and lower bounds for the limit.

Deken [5] showed that (L_n/n) converges almost surely to a random variable under a stationarity condition on the LLCS, and that if the two words are written independently of each other and the alphabet letters are equiprobable - an assumption untenable for biological data - the limit is a constant. Finally Steele [14] showed that if the vectors (X_i,Y_i) are I.I.D., $VarL_n = O(n)$ and proposed the replacement of the LLCS by other statistics in view of its intractability.

Suppose that $\underline{X}=(X_1,\ldots,X_m)$ and $\underline{Y}=(Y_1,\ldots,Y_n)$ with X_t and Y_t obtaining values in a finite alphabet $\{a_1,\ldots,a_s\}$, assume that $m \geq n$ and let $\{M_{ij}\}$ $i=1,\ldots,m$ $j=1,\ldots,n$ be the CC matrix of $\underline{X}$ and $\underline{Y}$. In the visual examination of a character matrix, in order to detect substrings common to both words, the investigator tilts the character matrix, aligns his axis of vision to the matrix entries $\{M_{i,i+k}\}_{i=1,\ldots,\min(m,n-k)}$ and searches for consecutive non-blank matrix entries along the matrix diagonals $\{M_{i,i+k}\}$.

In order to fix ideas we introduce some new nomenclature. If $\underline{X}$ and $\underline{Y}$ share in common a relatively long substring so that

$$X_u = Y_{u+k}, X_{u+1} = Y_{u+1+k}, \ldots, X_v = Y_{v+k} \tag{4-1}$$

for $1 \leq u \leq v \leq m$ and $1 \leq u+k \leq v+k \leq n$,

then the entries $M_{u,u+k}, M_{u+1,u+k+1} \cdots, M_{v,v+k}$ will be nonblank as shown in figures 4-2a and 4-2b and we shall say that words $\underline{X}$ and $\underline{Y}$ share in common a string of length v-u+1 lying along the diagonal at lag k, or, more briefly, that <u>X and Y share a common string at lag k</u>.

Let M_{ij} be the CC matrix for $\underline{X}$ and $\underline{Y}$ and suppose that $k \geq 0$. A long common string in (4-1) would cause the ratio of non-blank matrix entries to the total number of entries on the diagonal $\{M_{1,1+k}, \ldots, M_{m-k,n}\}$ to be higher than ratios on parallel diagonals of comparable length as indicated in figures 4-2b and 4-2c. We shall call the ratio of nonblank matrix entries on $\{M_{1,1+k}, \ldots, M_{m-k,n}\}$ to the total number of matrix entries along the diagonal (i.e. m-k), <u>the diagonal smear at lag k</u>.

For $m \geq n$, the process of diagonal smears can be written as:

$$D(k) = \frac{\sum_{j=1}^{n-k} \phi(X_j, Y_{j+k})}{n-k} \qquad \text{if } k \geq 0$$

$$D(k) = \frac{\sum_{j=1}^{\min(n,m+k)} \phi(X_{j-k}, Y_j)}{\min(n,m+k)} \qquad \text{if } k < 0 \tag{4-2}$$

for $\phi(.,.)$ defined in (3-6).

The process of diagonal smears is relevant in detecting common substrings among $\underline{X}$ and $\underline{Y}$ because a common substring would cause the diagonal smear at a lag specified by the string's position in the two words to be relatively high and, conversely, lags at which diagonal smears are high could signify the presence of a common substring.

"The proof of the pudding is in the eating." If the process D(.) proposed is of any value, it should pick up diagonals where they exist and indicate that there is nothing of interest where there are no diagonals. The performance of D(.) will be assessed on chorion proteins 292 and 18B which were examined visually in the development of the variety of character matrices in chapter 2. The cytochrome c protein of Tetrahymena pyriformis and the chorion 292 protein were chosen as a "control" pair because it was expected that they would share no similarity whatsoever as they play very different rôles in the lives of two distant organisms.

Figures 4-3a and 4-3b present the CC and the BNC1 character matrices for the control pair and illustrate that, as expected, the proteins of the control pair share no long strings in common. The longest common string is three letters long, while the longest string common in both proteins up to non consecutive mismatches is only four letters long. Figures 4-4a and 4-4b plot diagonal smears vs. lag for chorion proteins 292 and 18B and the control pair. For diagonals at highly positive or highly negative lags diagonal smears are computed for a small number of observations; this is the reason for which the variability of D_k is higher in the left and right tails of the plots than in the middle.

As illustrated from their BNC1 matrix on figure 4-5, the three most prominent strings common to the chorion proteins 292 and 18B lie along the diagonals at lags -12, -10 and 0, other prominent common strings lying, in order of diminishing prominence, on the diagonals at lags -15,-5, -20, 5 and -100. Table 4-1 lists the twenty-four largest

diagonal smears in decreasing order.

Table 4-1. Sorted diagonal smears for proteins 292 and 18B.

RANK	LAG	D.SMEAR	RANK	LAG	D.SMEAR
1	-12	.48	13	63	.21
2	0	.36	14	-114	.20
3	-100	.29	15	-129	.20
4	-98	.25	16	10	.20
5	5	.24	17	-83	.20
6	-24	.24	18	-29	.19
7	-10	.22	19	57	.19
8	67	.22	20	73	.19
9	-18	.22	21	-26	.19
10	-2	.21	22	94	.19
11	-5	.21	23	-20	.18
12	-96	.21	24	72	.18

It can be seen from table 4-1 that the diagonal smears at lags -12, -10, 0, -15, -5, -20, 5 and -100 (where prominent common strings lie) are the first, second, seventh, thirtieth, eleventh, twenty third, fifth and third largest. If there was a nonblank character along one of the diagonals of length two, its diagonal smear (.5) would be higher than any of the above. Clearly, it does not suffice to simply sort diagonal smears in decreasing order. The threshold above which diagonal smears should be considered as "significantly" high must depend on diagonal length.

Under the independence model, both the matrix smear and the diagonal smear estimate the same parameter. The matrix smear of the CC matrix for two words is computed from all blank and nonblank entries of the matrix. The diagonal smear estimates σ from the ratio of non-blank characters on the diagonal. Under the independence assumptions the matrix smear has been proven to be asymptotically normally distributed about the theoretical smear and the number of non-blank characters on the diagonal is binomially distributed. Hence, an upper confidence bound from

the matrix smear and a lower confidence bound for the same parameter from the binomial data at each diagonal may be computed for σ. Throughout the remainder of the chapter words will be assumed to be written independently within and between themselves.

Let U be a $1-\alpha_1$ upper confidence bound for σ computed from S through proposition 3-1. Let $\hat{V}$ be the m.l.e. of the asymptotic variance V given in equation (3-4). Then, clearly $\hat{V}$ converges in probability to V and

$$U = S + \frac{\sqrt{\hat{V}}\, Z_{1-\alpha_1}}{\sqrt{m}} + o_p(1/\sqrt{m}), \tag{4-3}$$

where $Z_{1-\alpha}$ is the $(1-\alpha_1)$ quantile of the standard normal distribution. For long words,

$$Pr(U>\sigma) \cong 1-\alpha_1, \tag{4-4}$$

Table 4-2 lists below the 90%, 95% and 99% asymptotic confidence intervals for σ computed from S by proposition 3-1a.

Table 4-2. Asymptotic confidence for the theoretical smears of CCM for protein pairs (292,18B) and (292, Cytochrome c).

$1-\alpha_1$	292, 18B	292, Cytochr. c
.90	(.11-.16)	(.06-.09)
.95	(.11-.16)	(.06-.09)
.99	(.10-.17)	(.05-.10).

The length of the confidence interval does not depend crucially on the confidence level up to the second decimal digit because the estimate of the asymptotic standard deviation of S in equation (3-2) is small.

Figures 4-6a and 4-6b plot D_k and the asymptotic two sided 95% confidence interval for σ for each of the protein pairs.

Let L_k be an exact $1-\alpha_2$ lower confidence bound for σ computed from

the binomial data along the diagonal of lag k. Suppose that the length of the matrix diagonal at lag k is N and that there are B nonblank entries on the diagonal. L_k is defined as:

$$L_k = \begin{cases} 0 & \text{if } B=0 \\ \text{the root of the equation } \sum_{i=B}^{N} \binom{N}{i} x^i (1-x)^{N-i} = \alpha_2 & \text{if } B>0, \end{cases} \tag{4-5}$$

and

$$Pr(L_k < \sigma) \geq 1-\alpha_2. \tag{4-6}$$

(See, for example, p. 181. of [1].)

What use is to be made of U and L_k? The hypothesis that $\sigma = \sigma_0$ is rejected at level α_2 in favour of the hypothesis $\sigma > \sigma_0$ if L_k exceeds σ_0. In our context no σ_0 is given to be tested; a $(1-\alpha_1)$ upper confidence bound may be set for the theoretical smear. It is then reasonable to suspect that when

$$U < L_k, \tag{4-7}$$

a string is common to the words $\underline{X}$ and $\underline{Y}$ at lag k, and expect that if $\underline{X}$ and $\underline{Y}$ share in common a long string, then inequality (4-7) will hold for a lag k specified by the position of the string in the two words. Hence to detect diagonals hosting long common strings, instead of sorting diagonal smears, we propose to compare L_k to U. Qualitatively, the $\{L_k\}$ relate to U as the $\{D_k\}$ to S; the presence of a long string common to the two words under examination raises L_k and has little effect on U. The advantage of L_k (vs. D_k) is that L_k take into account diagonal length and consequently the variability of L_k is smaller than that of D_k as can be seen by comparing plots of D_k and L_k. Figures 4-7a and 4-7b plot the 97.5% upper confidence bound U and the 97.5% lower confidence bound

L_k at each lag, for the two selected protein pairs.

The probability of the event at (4-7) is

$$\Pr(U < L_k) = \Pr(\sigma \leq U < L_k \text{ or } U < \sigma < L_k \text{ or } U < L_k \leq \sigma)$$

$$= \Pr(\sigma \leq U < L_k \text{ or } U < \sigma < L_k) + \Pr(U < \sigma < L_k \text{ or } U < L_k \leq \sigma) - \Pr(U < \sigma < L_k)$$

$$= \Pr(\sigma < L_k) + \Pr(U < \sigma) - \Pr(U < \sigma < L_k).$$

In view of (4-4) and (4-6), an upper bound for the event in (4-7) is:

$$\Pr(U < L_k) \leq \alpha_1 + \alpha_2. \qquad (4\text{-}8)$$

When (4-7) holds, we shall say that <u>the diagonal smear is significantly larger than the matrix smear at level $\alpha_1+\alpha_2$</u> .

As it can be seen from table 4-2, at $\alpha_1=.025$, U equals .16 for the pair of chorion proteins and .09 for the control pair. The lags at which diagonal smears are significantly larger than matrix smears for both protein pairs are listed below.

Table 4-3. Lags at which diagonal smears are significantly higher than the CC matrix smears of protein pairs (292,18B) and (292, Cytochr. c). $\alpha_1=.025$, $\alpha_2=.025$

292, 18B		292, Cyto c	
LAG	LCB	LAG	LCB
-12	.39	-	-
0	.28	-	-
5	.17	-	-

At $\alpha_1=\alpha_2=.025$, the proposed procedure detects the matrix diagonals on which the three most prominent strings common to 292 and 18B lie. No diagonal smears are significantly higher than the matrix smear at these levels, in the control pair.

To detect more diagonals one should either lower U or raise L_k, i.e. increase either α_1 or α_2. From table 4-2 it can be seen that (up to the second decimal point) the asymptotic 95% UCB for σ is .16.; the lags

of the diagonals at which the diagonal smear is significantly higher than the matrix smear for α_1=.05 and α_2=.025 are also given by table 4-3.

Figures 4-7c and 4-7d plot the same bounds of σ for α_1=.025 and α_2=.05. The lags at which diagonal smears are significantly higher than matrix smears at these levels are given in table 4-4.

Table 4-4. Lags at which diagonal smears are significantly higher than the CC matrix smears of protein pairs (292,18B) and (292, Cytochr. c). α_1=.025, α_2=.05

292, 18B		292, Cyt.c	
LAG	LCB	LAG	LCB
-100	.17	-72	.09
-24	.17		
-12	.40		
-10	.16		
0	.29		
5	.18		

Besides the diagonals already detected in table 4-3, the two next prominent strings in the BNC1 matrix of the chorion proteins are detected in table 4-4 and indication is given that a string common to both words might occur along the diagonal at lag -24. The BNC1 matrix of figure 4-5 indicates that the longest common string along this diagonal is the tetrapeptide AVAG. On the other hand, in the control pair and at the same levels, the diagonal smear at lag -72 is significantly larger than the matrix smear and the detection is void of any biological content. Hence, the control pair does not allow us to consider the tetrapeptide selected for the chorion pair as the realization of a legitimate signal.

It was desirable to derive a simultaneous confidence band for σ at each lag. As this has not been attained, a $1-\alpha_1$ upper confidence bound for σ from S and a $1-\alpha_2$ lower confidence bound for the same parameter from D_k are constructed and the lags of diagonals at which the diagonal

smears are significantly larger than the matrix smear are listed for further examination. To detect the common susbtrings in the data the investigator now focuses on the selected diagonals. A procedure to automate this detection will be proposed in chapter 5. In this chapter the detection is carried out by a visual examination along the diagonals of the BNC1 matrix.

For $\underline{X}=(X_1,\ldots,X_m)$ and $\underline{Y}=(Y_1,\ldots,Y_n)$ let $m\geq n$ and $L\geq 0$ without loss of generality and let $\mathbf{M} = \{M_{ij}\}$ be a character matrix at the disposal of the investigator. The diagonal of $\mathbf{M}$ at lag L aligns the substrings

$$\begin{aligned} &X_1 \ ,\ldots,X_{n-L} \\ &Y_{1+L},\ldots,Y_n. \end{aligned} \tag{4-9}$$

If

$$M_{a,a+L},\ldots,M_{b,b+L} \tag{4-10}$$

is the prominent substring of mostly non-blank character entries on the diagonal of $\mathbf{M}$ at lag L, then

$$\begin{aligned} &X_a \ ,\ldots X_b \\ &Y_{a+L},\ldots,Y_{b+L} \end{aligned} \tag{4-12}$$

are the most similar substrings of $\underline{X}$ and $\underline{Y}$. The substrings of (4-10) can be thought of as two realizations of a signal in $\underline{X}$ and $\underline{Y}$; the mismatches between X_i and Y_i in (4-12) caused by the imposition of noise on the signal.

How good is proposed procedure? There are two types of error that the procedure may commit and which, following the use of the terms in the statistical literature, we call type I and type II errors.

A type I error occurs when no signal is present in both words and the procedure comes up with some diagonal smear significantly larger than

the matrix smear. A type I error may be thought of as a "false alarm".

No type I error was committed when the proposed procedure was applied to the control pair at α_1=.025 and α_2=.025. A false alarm was given for the same pair, at α_1=.025 and α_2=.05; one out of the 242 diagonal smears was significantly larger than the matrix smear. If $\{L_k-U\}$ were I.I.D. and the upper bound of $Pr(L_k>U)$ in inequality (4-8), was attained, we would expect that diagonal smears would be significantly higher than S at approximately 12 and 18 lags for the two sets of levels chosen. (Because .05*242=12.1 and .075*242=18.1.) The discrepancy between the observed and the expected is striking and can be attributed to two factors: $\alpha_1+\alpha_2$ is only an upper bound to $Pr\{U<L_k\}$ and $\{L_k-U\}$ are not independent. False alarms suggesting that very few out of hundreds of diagonal smears are significantly high (in our case 1 out of 242) are painless; in requesting the investigator to focus on a few diagonals, the proposed procedure reduces drastically the volume of work involved in the visual examination of the data .

A <u>type II error</u> arises when a string is common to the two words but it is not long enough to cause the diagonal smear at the lag specified by its position in the words, to become significantly larger than the matrix smear. While the occurrence of a type I error is rather painless, a type II error is a serious one.

The detection of common strings by comparing U to L_k for each diagonal was developed while examining chorion proteins 292 and 18B. The proposed procedure is now applied to the proteins encoded by the Balbiani ring genes which are denoted by BR1, BR2 and BRC and presented on figure 4-8a. Figure 4-8b lists the proteins products of the Balbiani

ring genes which will be called BR1, BR2 and BRC proteins. Figure 4-9 illustrates the BNC1 character matrix for proteins BR1 and BR2, underlines the strings most prominently common to the BR1 and BR2 proteins. The underlined strings lie on matrix diagonals at lags -173, -105, -91, -31, -23, -9, 51, 59 and 133. The underlined strings suggest that there are extensive internal repeats within each of BR1 and BR2 proteins; the repeats are illustrated in the BNC1 character matrices for the proteins on figures 4-10a and 4-10b. For the BR1 and BR2 proteins, S=.120. Asymptotic two-sided confidence intervals of σ at different levels, computed from proposition 3-1a, are presented in table 4-5 below.

Table 4-5. Two-sided confidence interval for σ from the CC matrix smear of the BR1 and BR2 proteins.

$(1-\alpha_1)$	Confidence Interval
.90	(.106-.134)
.95	(.104-.136)
.99	(.099-.141)

The process of diagonal smears and the 95% asymptotic confidence interval for σ are plotted on figure 4-11. Figure 4-12a plots U and L_k for α_1=.025, α_2=.01. As can be seen from table 4-5, the 97.5% UCB for σ from S is .136. The lags at which $\{L_k>U\}$ for α_1=.025 and α_2=.01 are given in table 4-6 below.

Table 4-6. Lags at which diagonal smears for the CC matrix of the BR1 and BR2 proteins are significantly higher than matrix smear. α_1=.025, α_2=.01.

LAG	LCB
-173	.366
-105	.197
- 91	.201
- 31	.166
- 23	.181
51	.212

If the strings detected visually and underlined on figure 4-9 are regarded as nine legitimate signals, table 4-6 indicates that at α_1=.025, α_2=.01 the proposed procedure commits no type I errors; it selects what appear to be the strongest six out of the nine signals on the BNC1 matrix of figure 4-9. To obtain the remaining signals - and at the risk of the occurrence of type I errors one must increase α_1 or α_2. For α_1 as large as .05, U= .134. At α_1=.05, α_2=.01, the procedure will still come up only with the smears of table 4-6 as significant; it is rather stable for fixed α_2. Figure 4-12b plots U and L_k for α_1=.025 and α_2=.025. The lags at which diagonal smears are higher than the matrix smear are given by table 4-7 below.

Table 4-7. Lags at which diagonal smears for the CC matrix of the BR1 and BR2 proteins smears are significantly higher than matrix smear. α_1=.025, α_2=.025

LAG	LCB
-173	.395
-105	.210
- 91	.214
- 31	.176
- 23	.192
51	.226
59	.139
143	.150
146	.139
152	.152

At α_1=.025 and α_2=.025, four more lags - besides the lags listed in table 4-6 - are selected : 59, 143, 146 and 152. Of those, the first one was detected after a visual examination of the BNC1 matrix on figure 4-9. No false alarms are given at the three remaining lags; strings are common to the BR1 and BR2 proteins, but they were not as prominent in their BNC1 character matrix to be picked up in the initial visual examination of the matrix. Finally the relatively short strings

underlined at diagonals of lags -9 and 133 should be considered as cases of type II errors when the procedure is operated on the BR1 and BR2 proteins at α_1=.025 and α_2=.025.

The strings underlined on the diagonals at lags 133 and -9 are too short to cause the corresponding diagonal smears to be significantly larger than the matrix smear. However, had protein BR1 been investigated for internal repeats, the two strings could have been detected from the strings underlined at lags 51 and -91 (at which as shown in tables 4-6 and 4-7 the diagonal smear is significantly higher than the matrix smear for both choices of α_1 and α_2) and the type II errors would have been eliminated.

For notational convenience denote the BR2 and BR1 proteins by $\underline{X}$ and $\underline{Y}$ respectively. The substring underlined on the diagonal at lag 133 aligns the octapeptide

$$X_{27} \ \ldots \ X_{34}$$

to $$Y_{160} \ \ldots \ Y_{168}. \qquad (4\text{-}12)$$

The substring underlined on diagonal at lag 51 aligns

$$X_{27} \ \ldots \ X_{63} \qquad (4\text{-}13)$$

to $$Y_{78} \ \ldots \ Y_{114}.$$

The most prominent diagonal of the BNC1 matrix for the BR1 protein on figure 4-10a (except for the trivial diagonal at lag 0) indicates that the substring $Y_1 \ldots Y_{82}$ is duplicated (exactly, with no mismatches) in $Y_{83} \ldots Y_{164}$. The repeat unit is partially triplicated in $Y_{165} \ldots Y_{16}$. With the understanding that entries in the same column are mostly identical, the repeat structure of BR1 may be summarized as:

$$Y_1 \ \ldots \ Y_4 \ \ldots \ Y_{78} \ \ldots \ Y_{82}$$

$$Y_{83} \cdots Y_{86} \cdots Y_{160} \cdots Y_{164} \qquad (4\text{-}14)$$

$$Y_{165} \cdots Y_{168}$$

Hence the common string of (4-12) can be inferred from those of (4-13) and (4-14) which are long enough to be detected. This suggests that before the procedure is applied to two different words it should be applied to each word for an investigation of internal repeats.

The proposed procedure depends on the parameters α_1 and α_2. It is desirable that the diagonals selected by the procedure be stable when the chosen levels α_1 and α_2 are slightly perturbed. The procedure depends on α_1 only through the quantile Z_{1-a} in U of equation (4-3). When $\alpha_1' \geq \alpha_1$, the set of selected diagonals at α_1' and α_2 includes the set of diagonals selected at α_1 and α_2. Table 4-8 sorts the twenty largest L_k for the two different values of α_2 at which the BR1 and BR2 proteins were examined.

Table 4-8. The twenty largest L_k for BR1 and BR2 proteins.

α_2=.01		α_2=.025	
LAG	LCB	LAG	LCB
-173	.37	-173	.39
51	.21	51	.23
-91	.20	-91	.21
-105	.20	-105	.21
-23	.18	-23	.19
-31	.17	-31	.18
59	.13	152	.15
143	.13	143	.15
152	.13	59	.14
16	.12	146	14
-19	.12	16	.13
146	.12	-19	.13
13	.11	149	.13
-93	.11	133	.12
-11	.11	13	.12
90	.11	90	.12
42	.11	-93	.12
133	.11	42	.12
-52	.11	-11	.12
87	.11	-161	.12

The table indicates that for both values of α_2, the six largest lower confidence bounds occur at the same diagonals. Eighteen out of the twenty lags listed for each value of α_2 overlap. The two non-overlapping lags for each value of α_2 being the lags for the twenty-first and twenty-second largest L_k at the other value of α_2. The proposed procedure possesses a desirable stability for α_2.

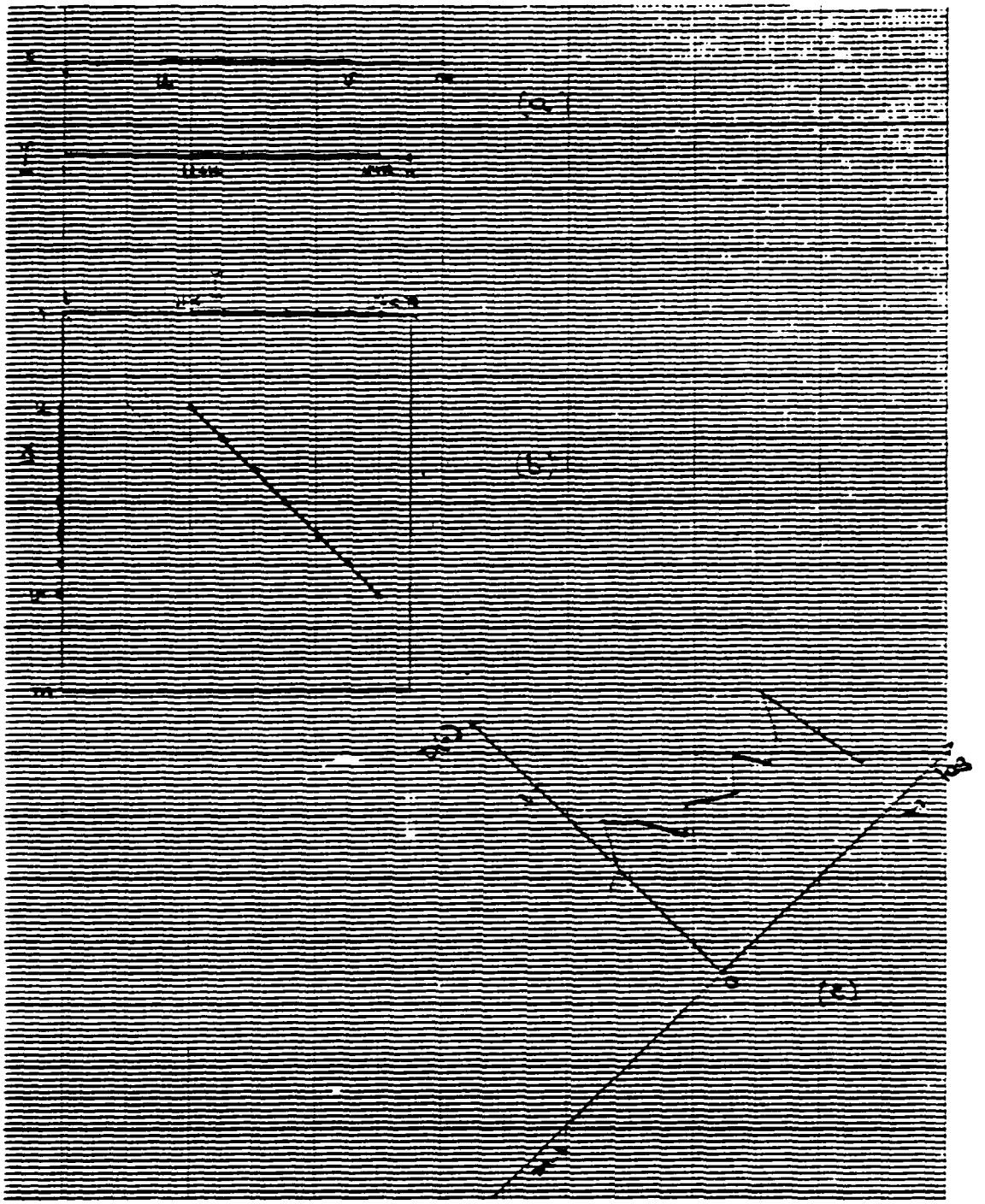


Fig. 4-2. (a) Words $\underline{X}$ and $\underline{Y}$ share in common the substring underlined at lag k. (b) A substring common to $\underline{X}$ and $\underline{Y}$ shows up in the CC matrix of $\underline{X}$ and $\underline{Y}$ as a diagonal as substring of nonblank matrix entries along the diagonal of lag k. (c) The diagonal smear plot associated with the CC matrix of $\underline{X}$ and $\underline{Y}$.

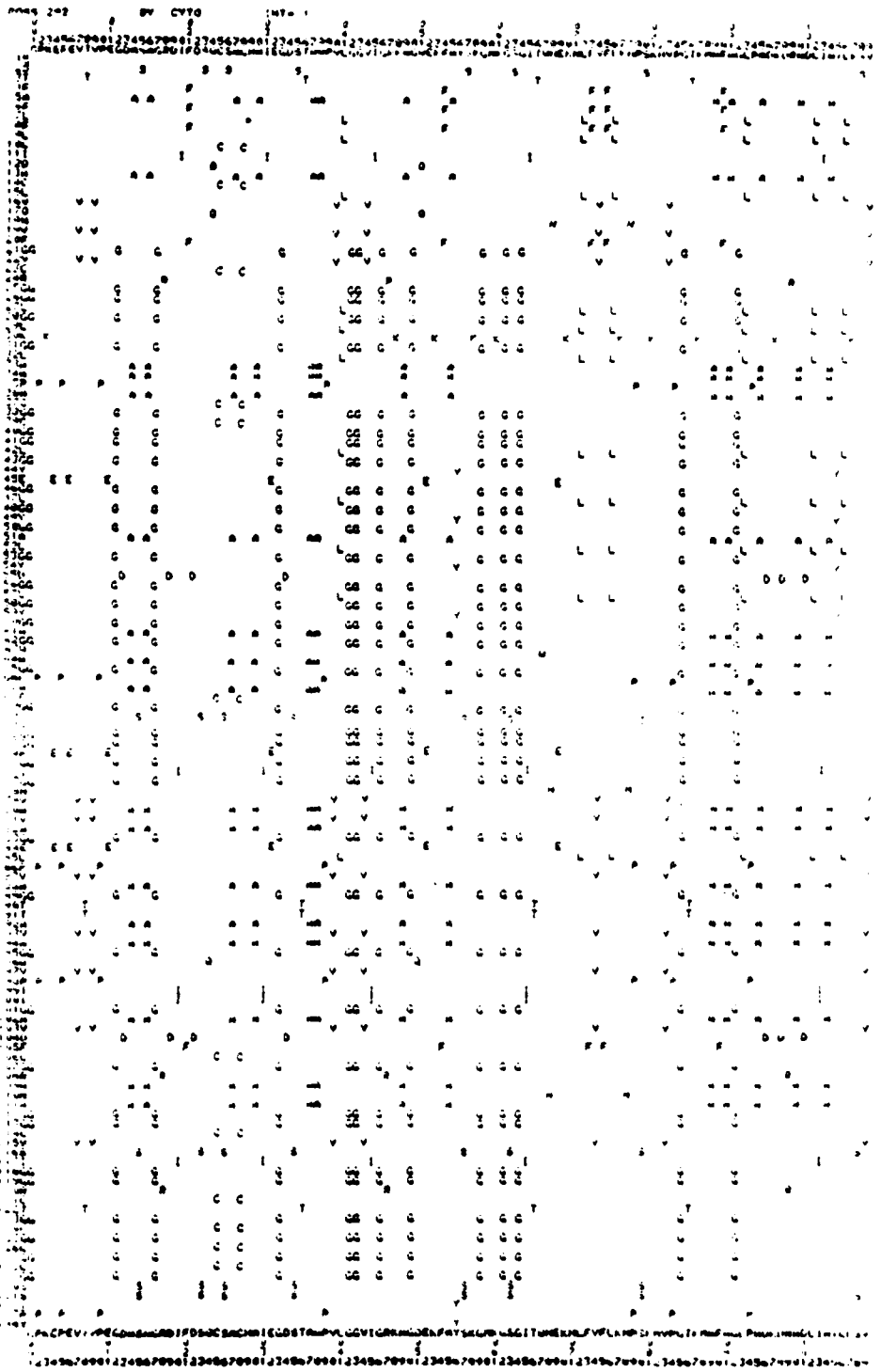


Fig. 4-3a. CC matrix for chorion 292 and cytochrome c proteins.

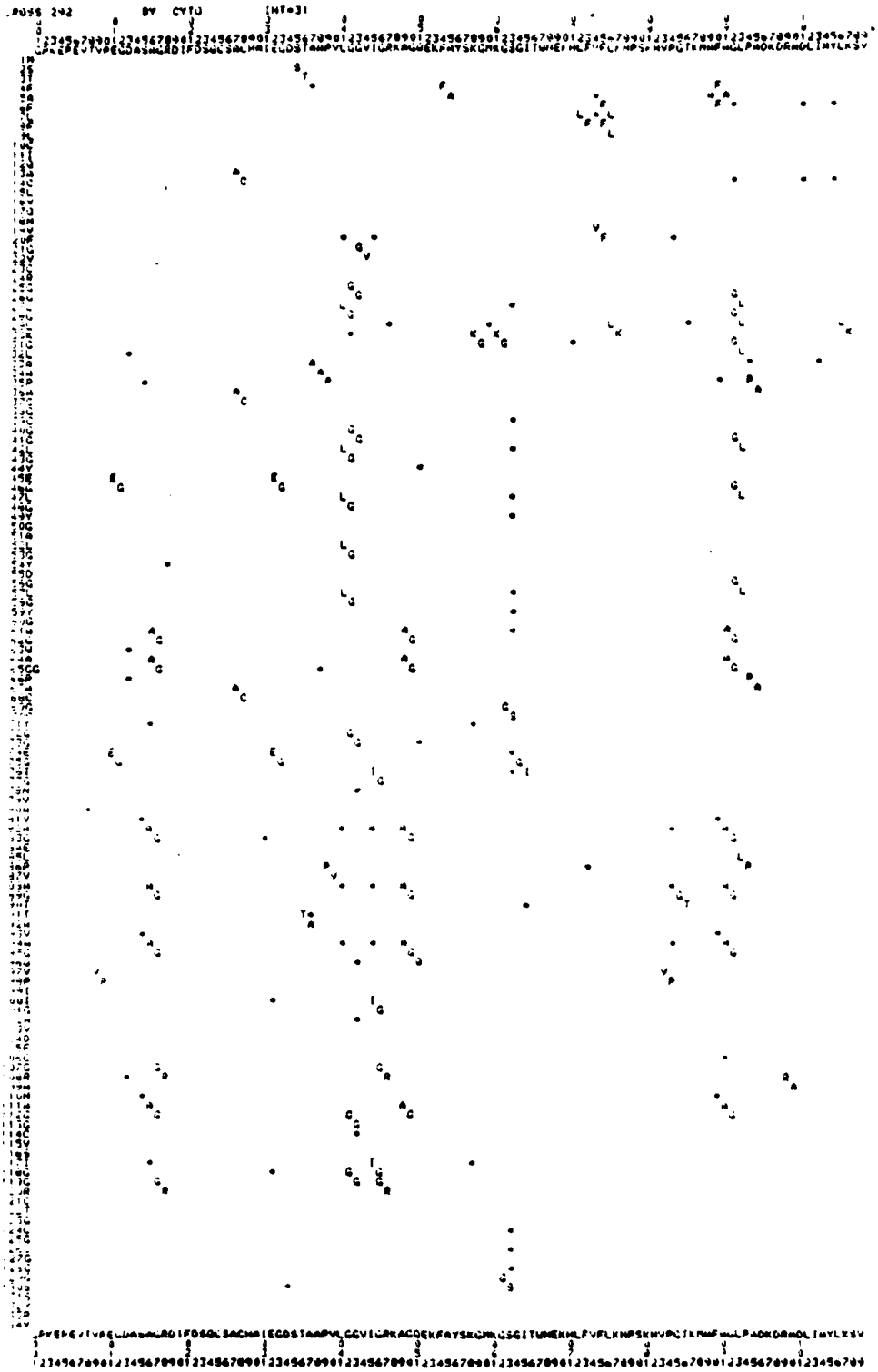


Fig. 4-3b. DMC1 matrix for chorion 292 and cytochrome c proteins.

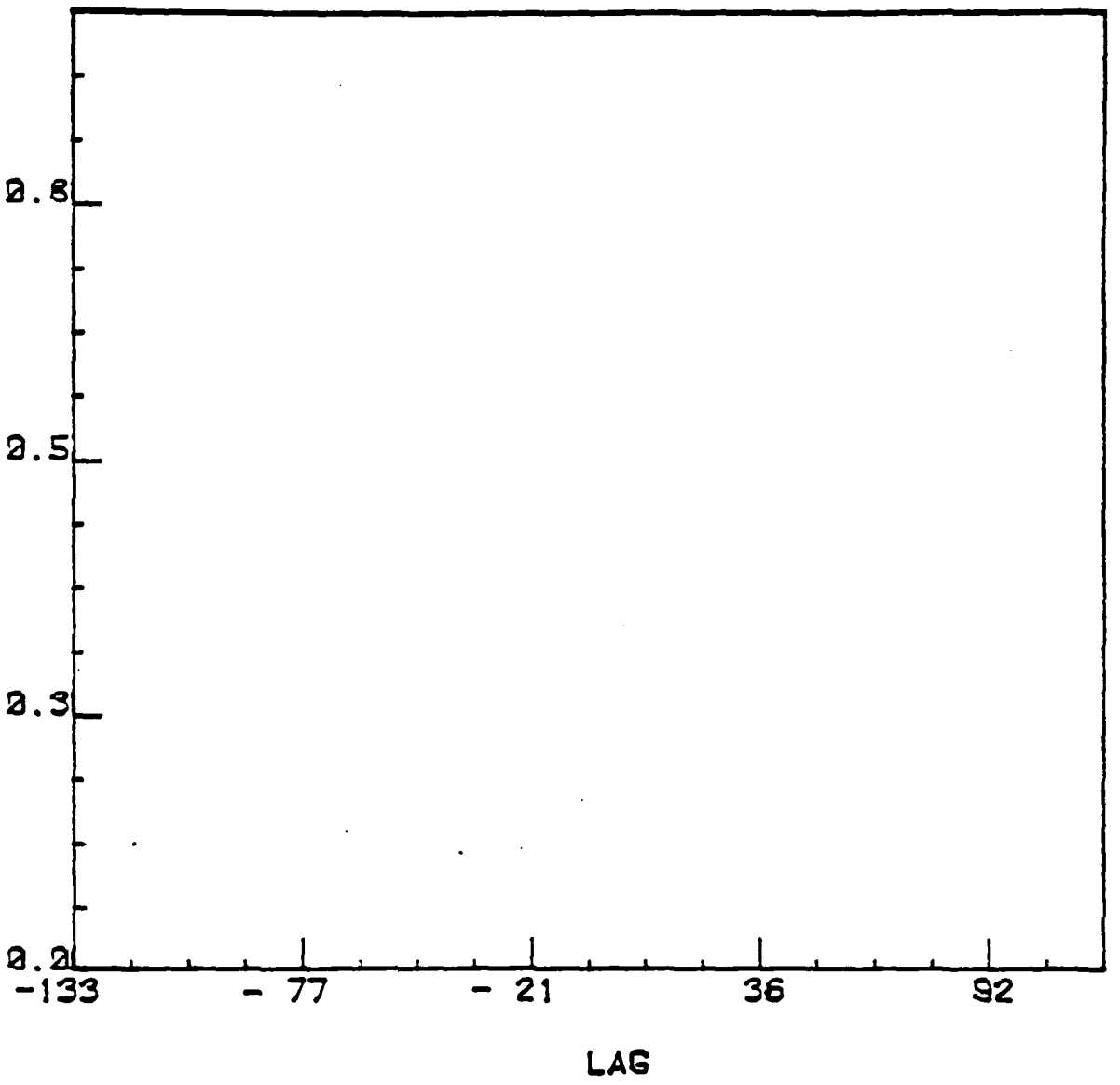


Fig. 4-4a. Diagonal smear of CCM plotted vs. lag for chorion proteins 292 and 18R.

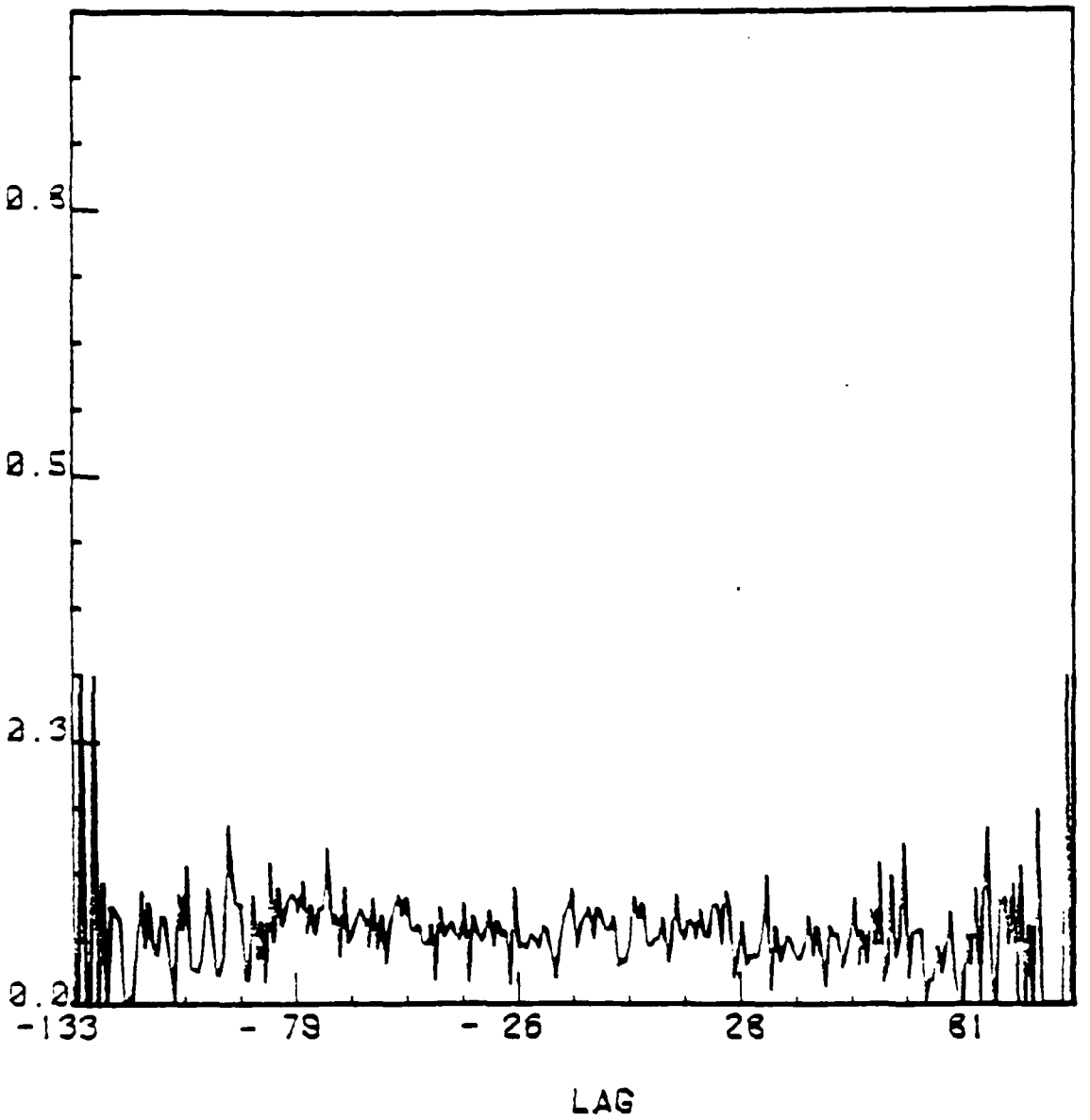


Fig. 4-4b. Diagonal smear of CCV plotted vs. lag for chorion 287 and cytochrome c protein.

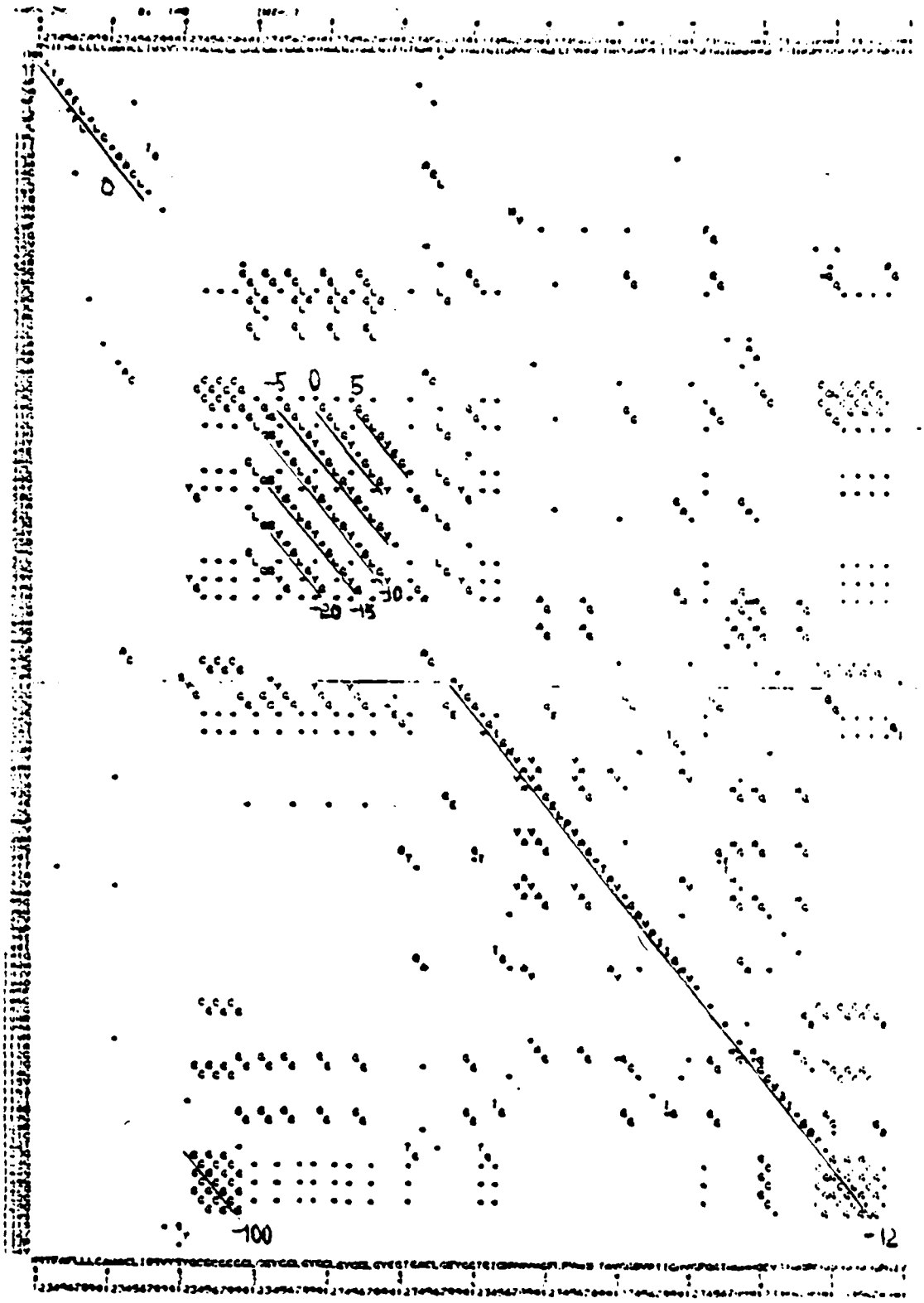


Fig. 4-5. BNC1 matrix for proteins 292 and 18B. Prominent common strings are underlined and the lags of their diagonals are indicated.

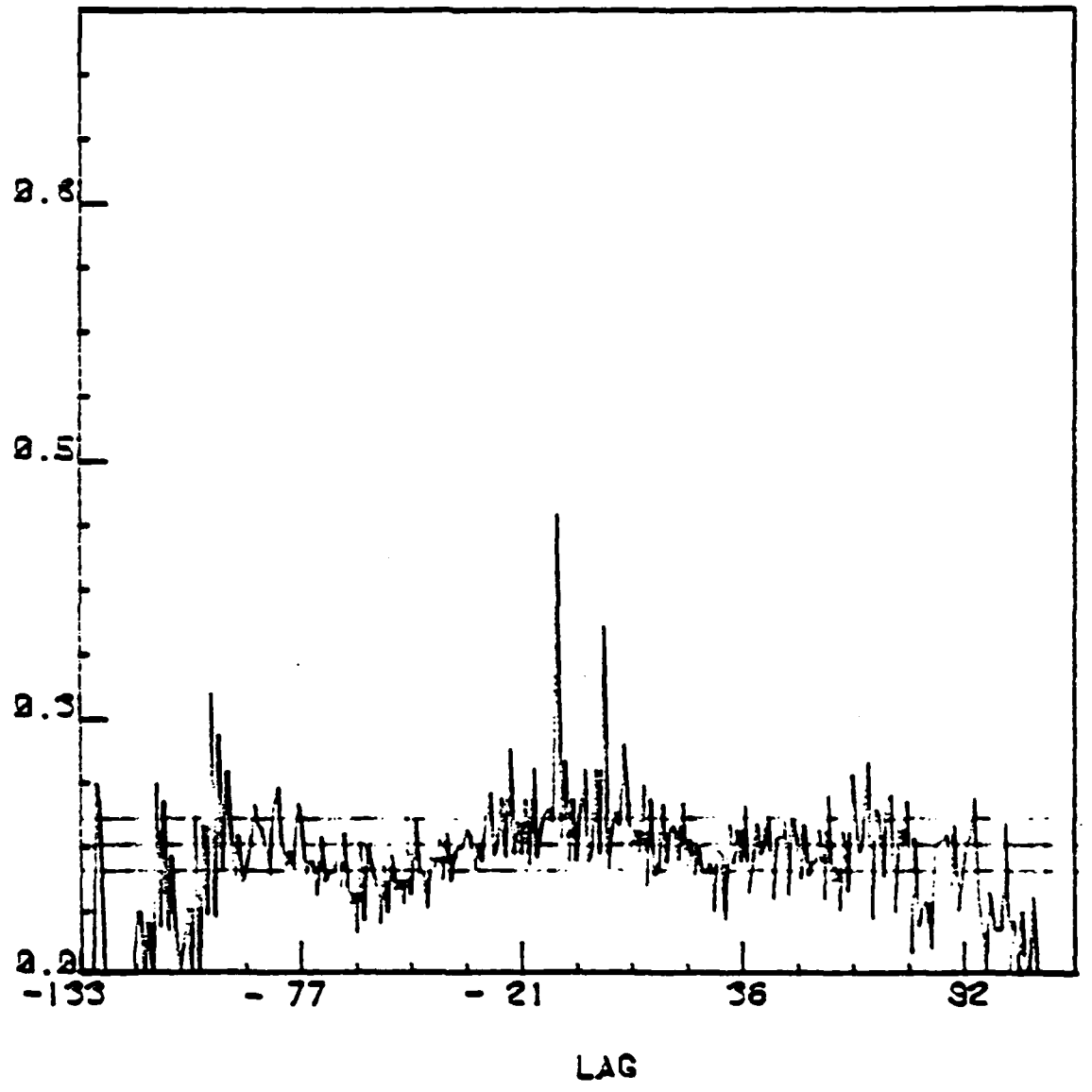


Fig. 4-6a. Diagonal smears for CCH of chorion proteins 292 and 132 and 95% asymptotic confidence interval for σ at each lag.

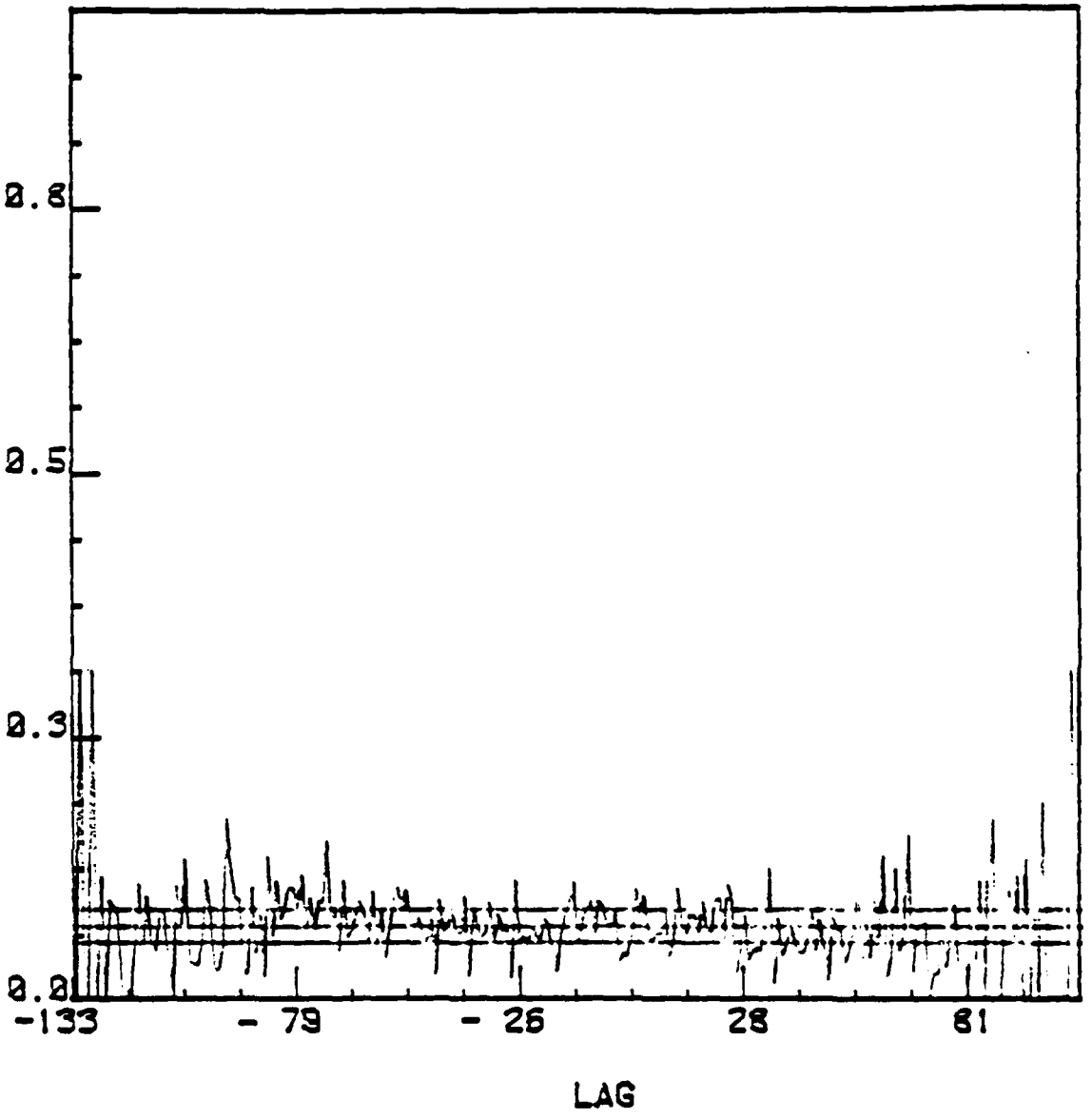


Fig. 4-6b. Diagonal smears for CCH of chorion 292 and cytochrome c proteins and 95% asymptotic confidence interval for σ at each lag.

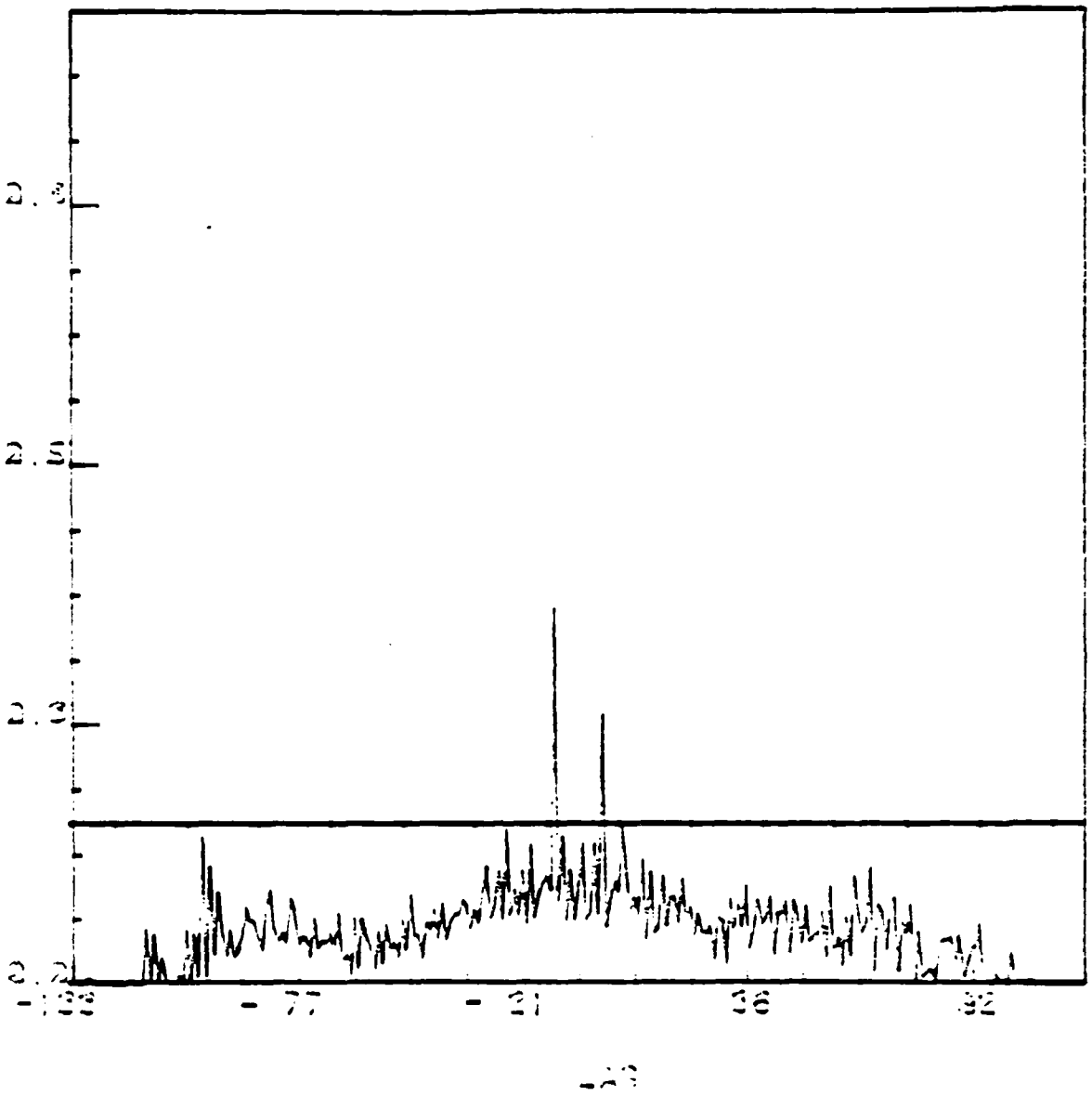


Fig. 4-7a. 97.5% Lower confidence bound of σ from diagonal smear and 97.5% upper confidence bound of σ from S. Chorion proteins 292 and 18B.

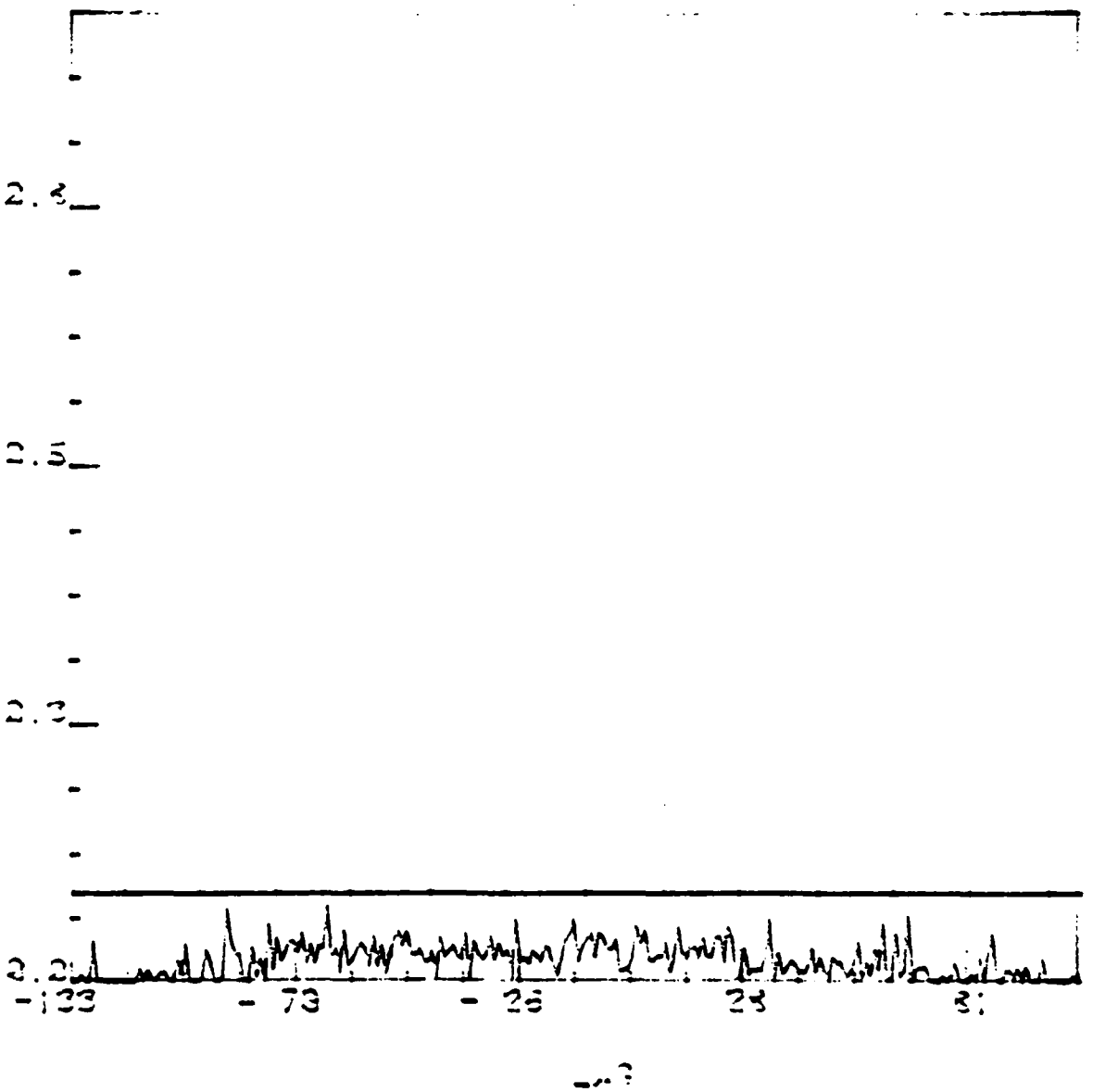


Fig. 4-7b. 97.5% Lower confidence bound of σ from diagonal smear and 97.5% upper confidence bound of σ from S. Control protein pair.

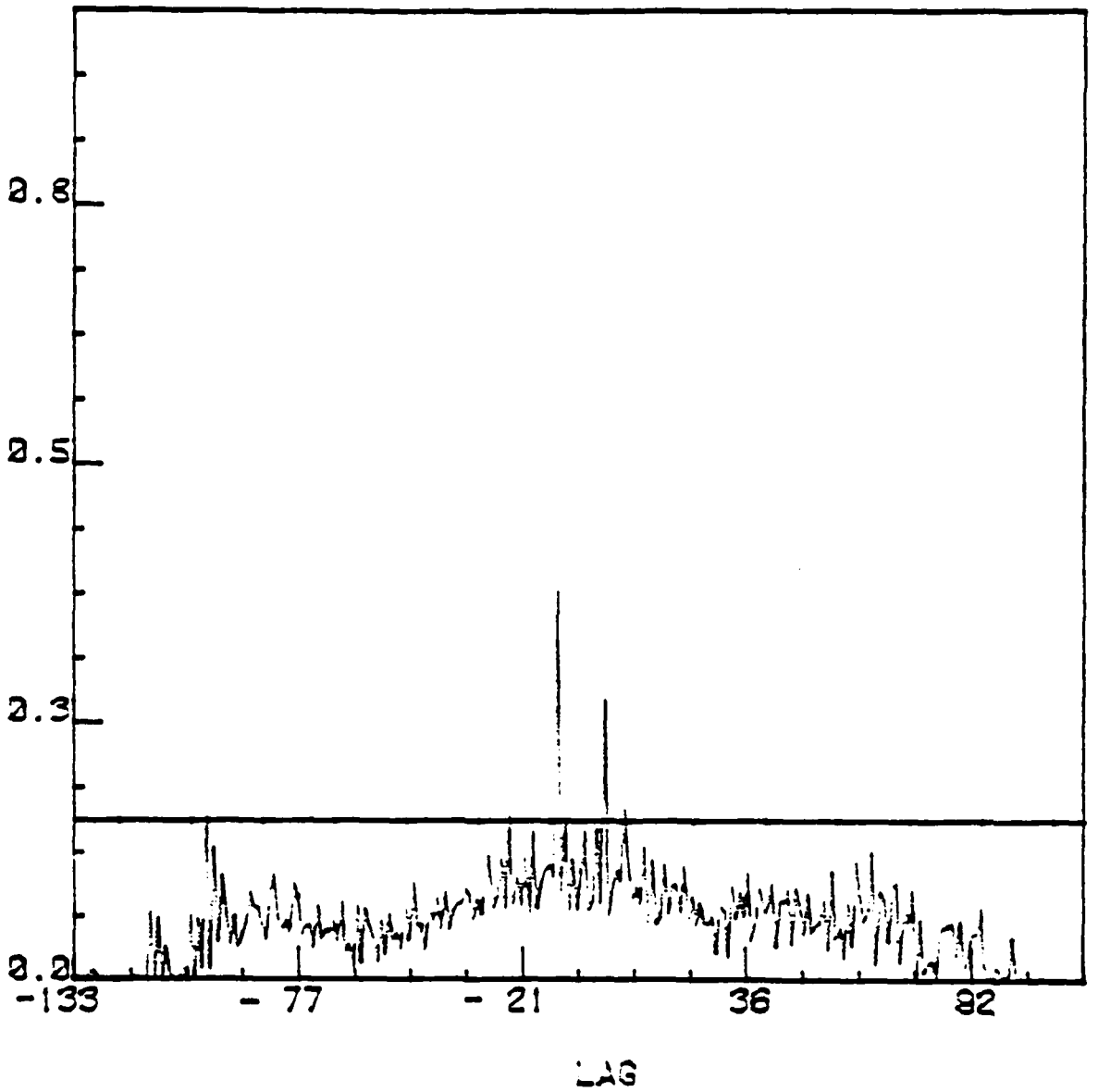


Fig. 4-7c. 95% Lower confidence bound of σ from diagonal smear and 97.5% upper confidence bound of σ from S. Chorion proteins 292 and 18B.

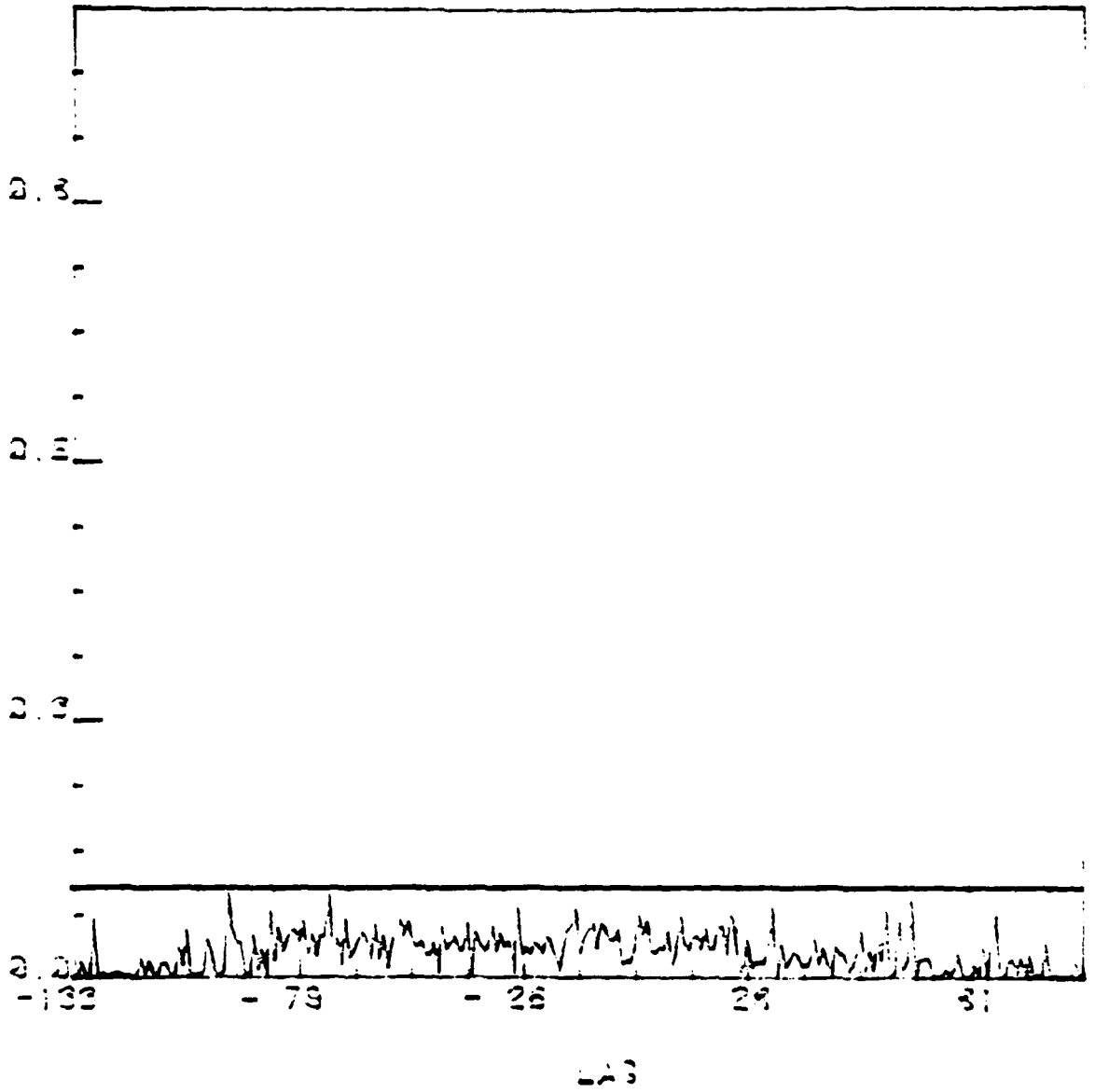


Fig. 4-7d. 95% Lower confidence bound of σ from diagonal smear and 97.5% upper confidence bound of σ from S. Control protein pair

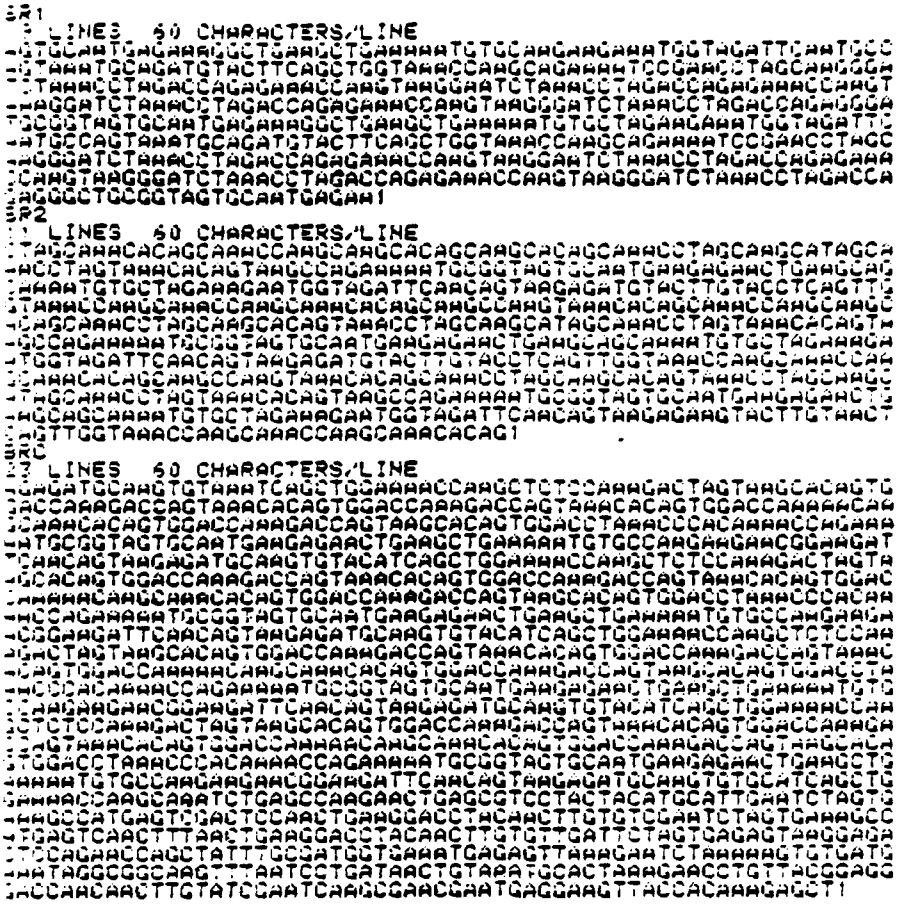


Fig. 4-8a. DNA sequences for Balbiani ring genes BR1, BR2, BRC.

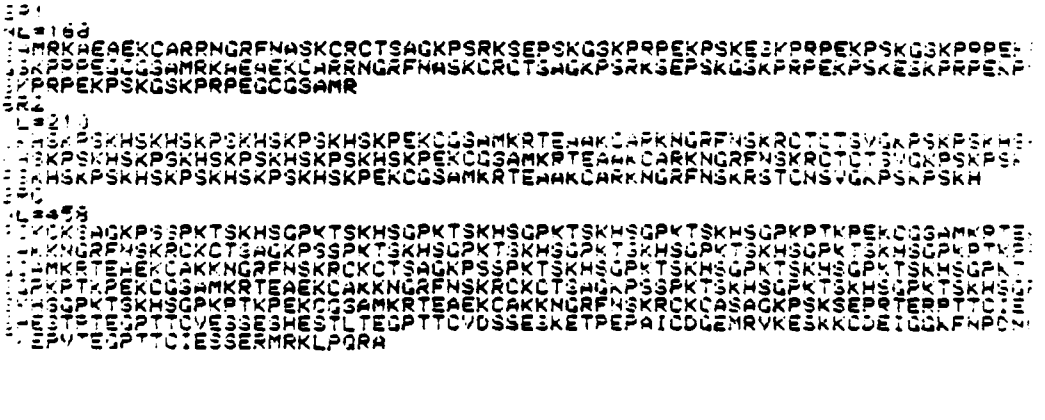


Fig. 4-8b. BR1, BR2, BRC proteins.

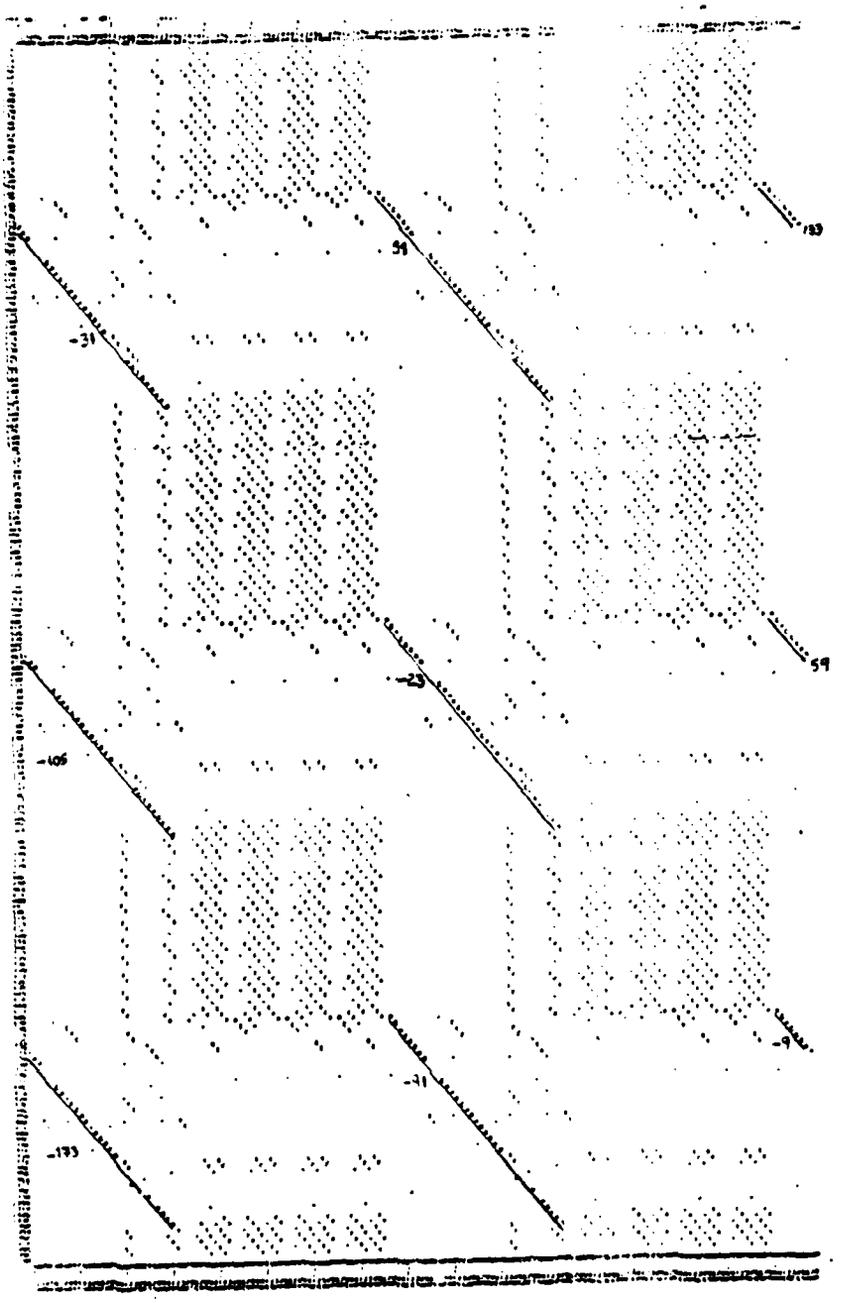


Fig. 4-9. BNC1 matrix for BR2 and BR1 proteins. The most prominent strings are underlined and the lags of their diagonals are indicated.

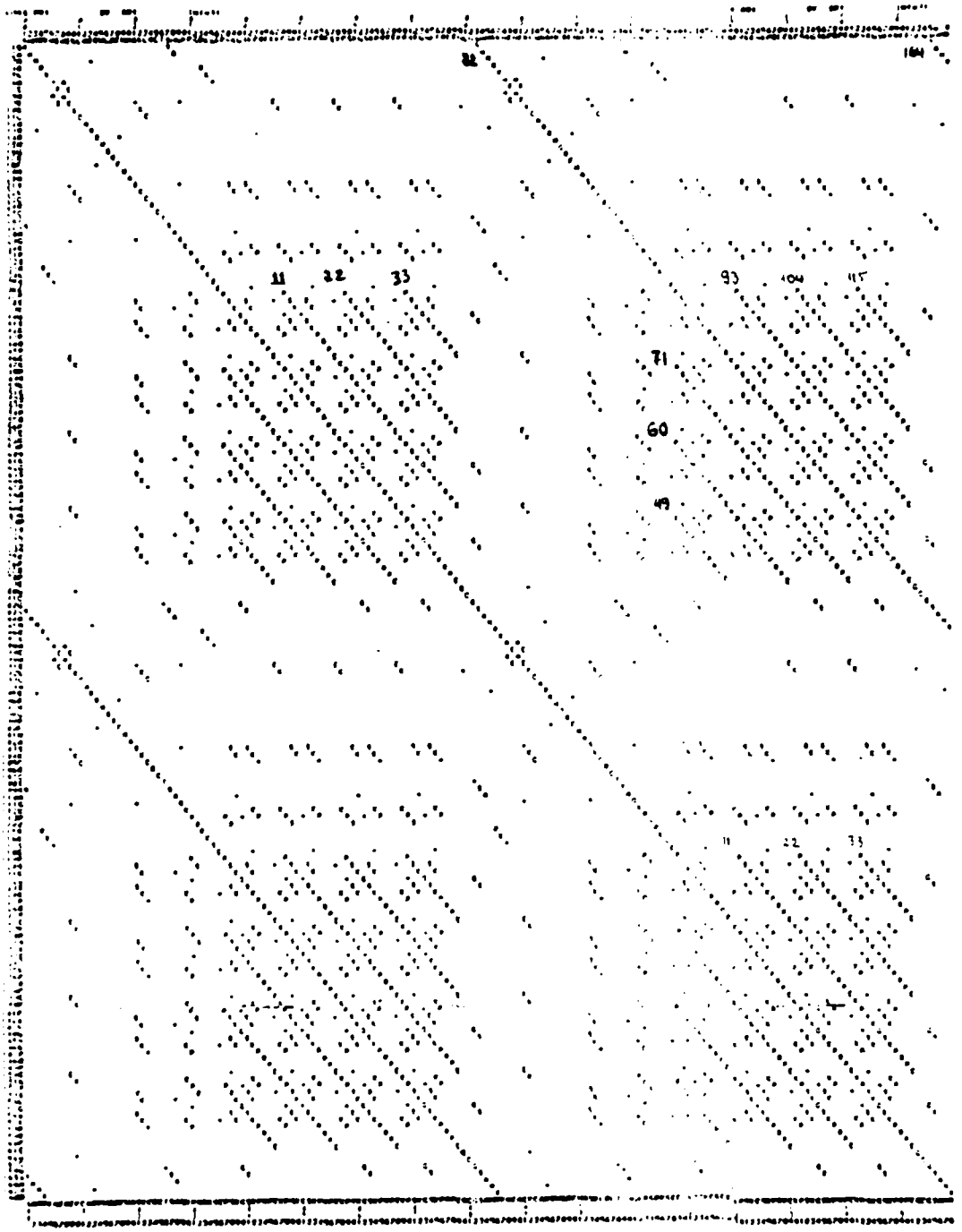


Fig. 4-10a. DNC1 matrix for HR1 protein.

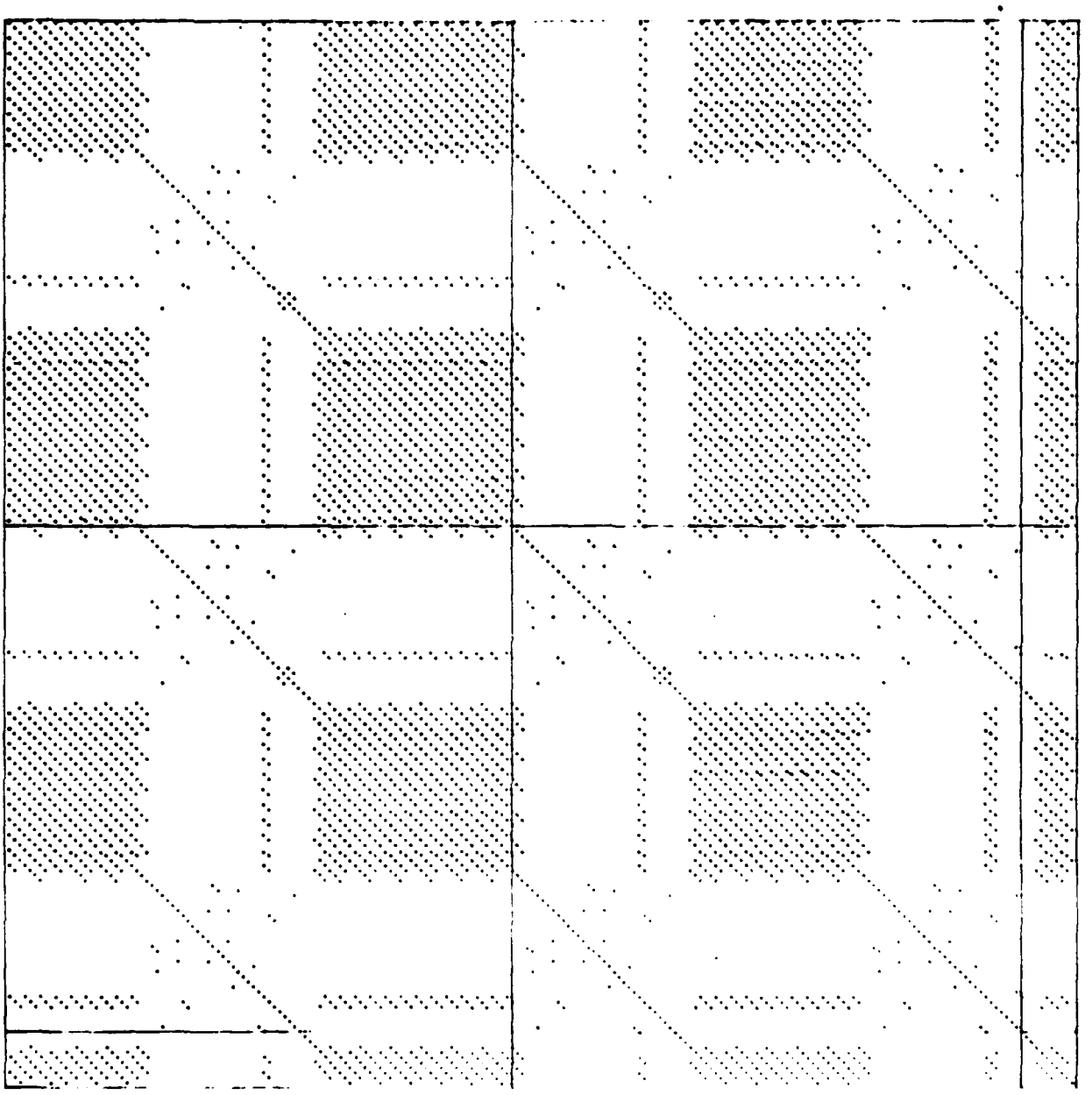


Fig. 4-10b. BNC1 matrix for BR2 protein. Solid lines are drawn every hundred amino acids.

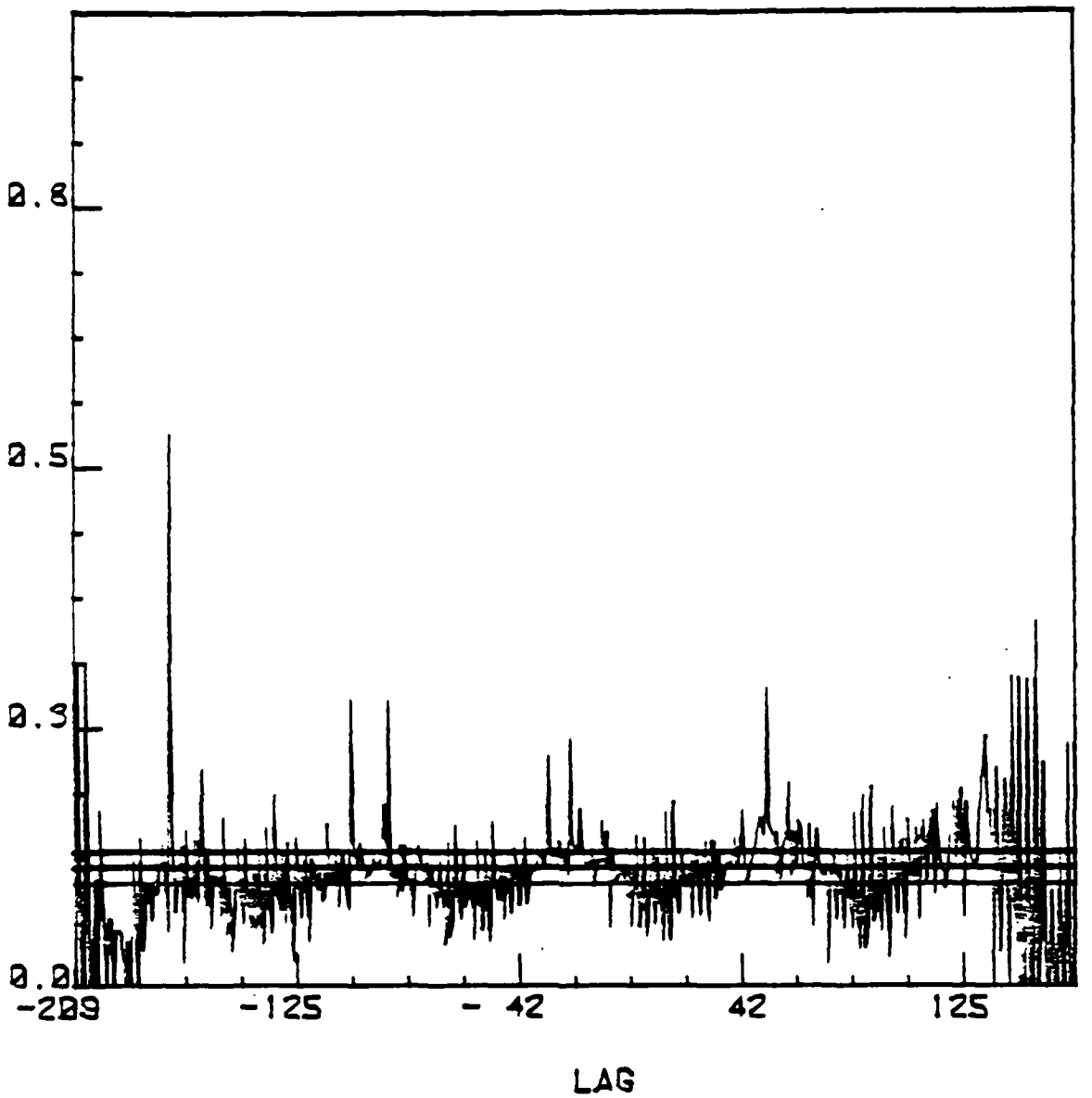


Fig. 4-11. Diagonal smears for CCF of P02 and P01 proteins and 95% asymptotic confidence interval for σ.

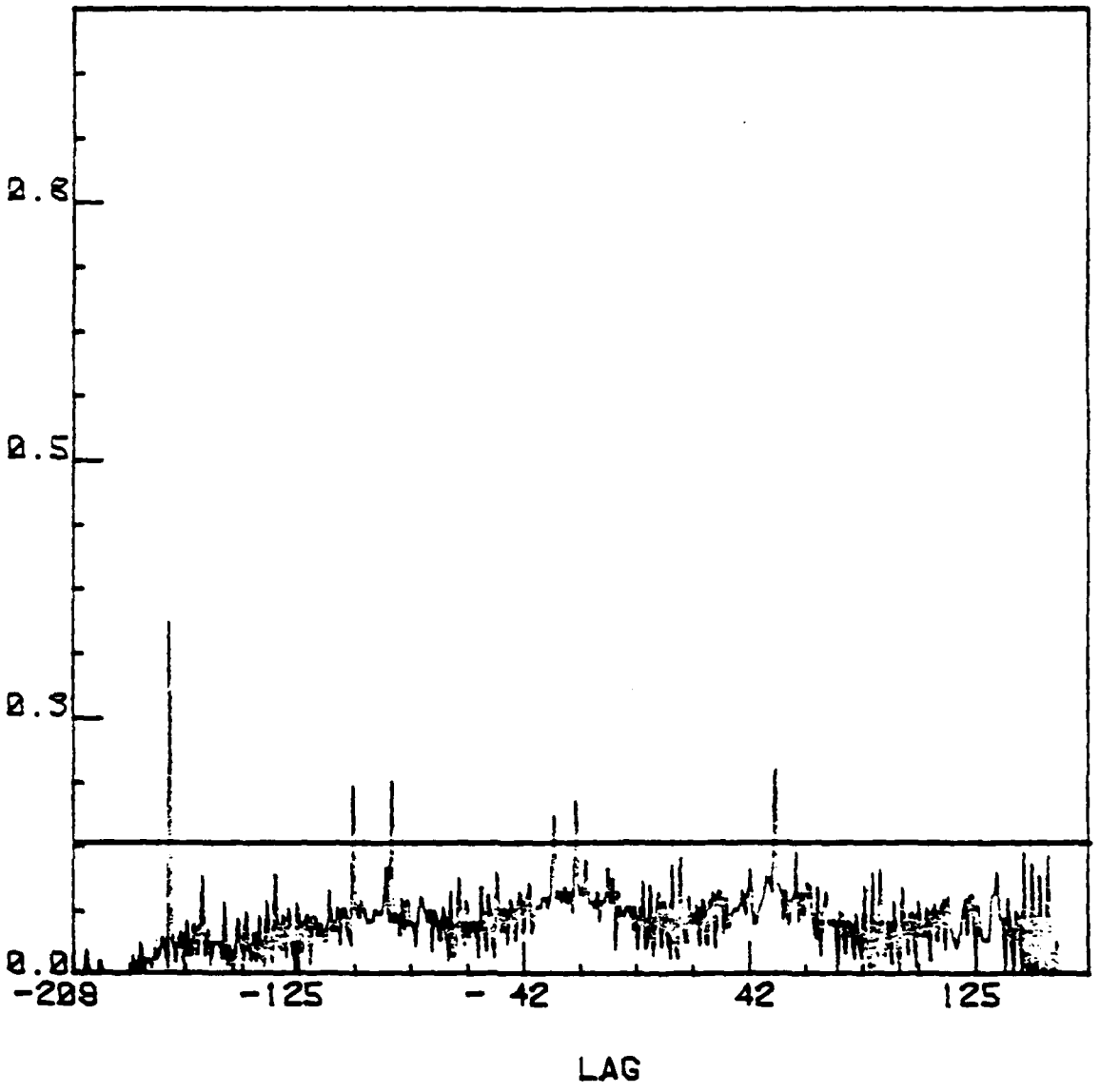


Fig. 4-12a 99% lower confidence bound of σ from diagonal smear and 97.5% upper confidence bound of σ from f for DR2 and DR1 proteins.

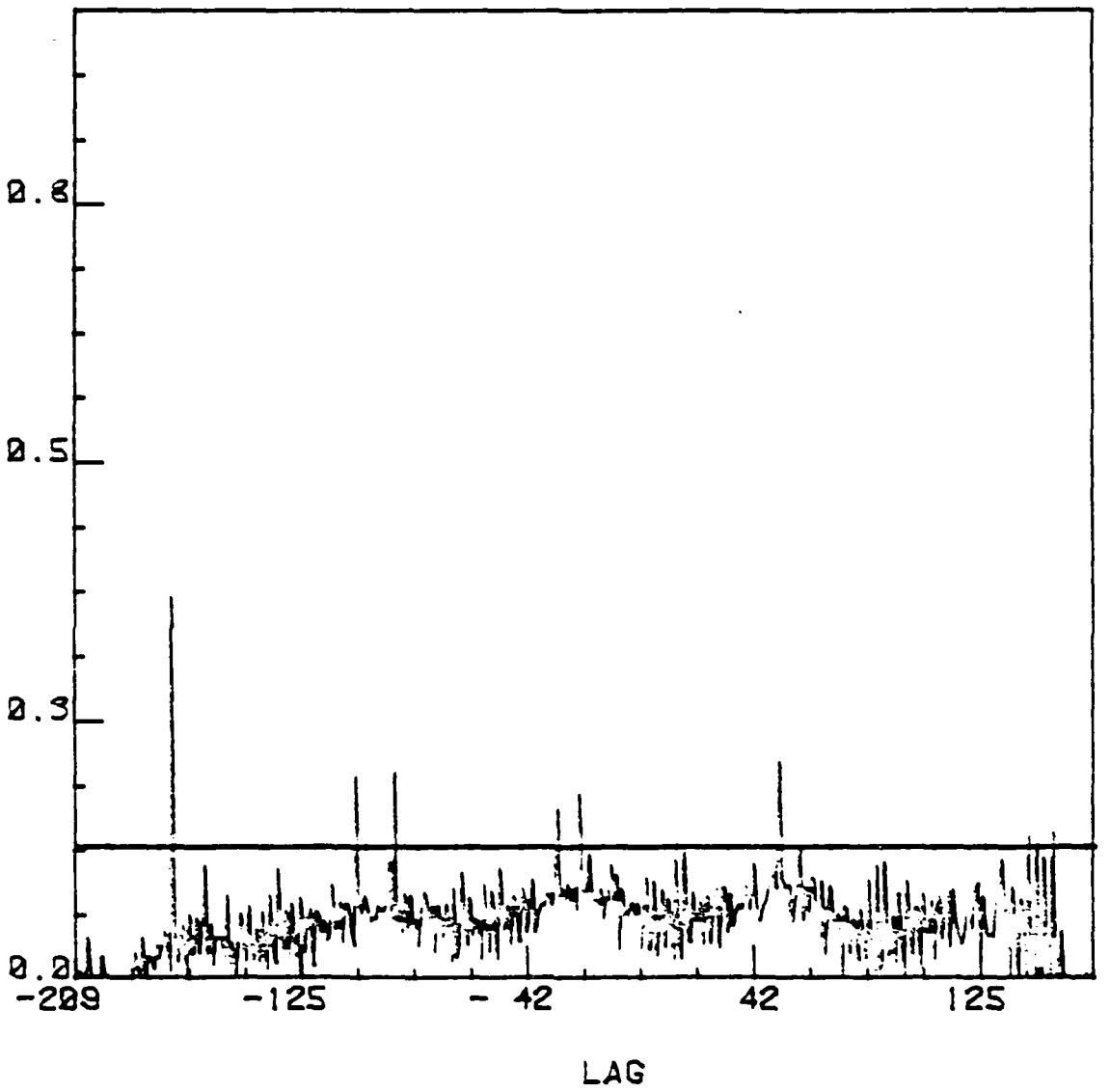


Fig. 4-12b 97.5% lower confidence bound of σ from diagonal smear and 97.5% upper confidence bound of σ from S for DR2 and DR1 proteins.

5. AUTOMATED DETECTION OF SIGNALS WITHIN TWO WORDS.

Chapter 4 introduced a procedure which automates partially the visual examination of character matrices of two words by focusing on the matrix diagonals for which the diagonal smear is significantly higher than the matrix smear. The detection of the common strings lying on the diagonals selected was carried out visually in chapter 4. This chapter proposes another procedure to automate the identification of the string most prominently shared in common by the two words under comparison. When a string will be referred to as shared in common by words $\underline{X}$ and $\underline{Y}$, it will be understood that a substring of $\underline{X}$ will be identical to a substring of $\underline{Y}$ except for a few occasional mismatches.

The proposed procedure is applied on matches and mismatches between substrings of the two words under comparison at all possible lags without taking into account the nature of the particular matches and mismatches. A procedure assigning weights to the latter is presumably more powerful than the one proposed here.

Suppose that the diagonal at lag L in the CC matrix of words $\underline{X}=(X_1,\ldots,X_m)$ and $\underline{Y}=(Y_1,\ldots,Y_n)$ is of length N. For notational convenience we denote the diagonal entries as

$$Z_1,\ldots,Z_N. \tag{5-1}$$

For the diagonal at lag L, $Z_i = \phi(X_i,Y_{i+L})$, ϕ defined by equation (3-6). In this chapter nonblank and blank matrix entries will be denoted by 1 and 0 and will be also called <u>successes</u> or <u>matches</u> and <u>errors</u> or <u>mismatches</u>. Independence between and within $\underline{X}$ and $\underline{Y}$ is assumed throughout the chapter. Under this assumption Z_i are I.I.D and the probability that Z_i is a nonblank character equals the theoretical smear of equation (2-3)

which in this chapter is denoted by p. (Instead of σ used in chapters 2, 3, and 4.) A string shared in common by $\underline{X}$ and $\underline{Y}$ will be called a signal. A signal at lag L will show up as a substring of (5-1) with a few occasional errors. Since only a few occasional mismatches are allowed in the realizations of the signal in the two words, detecting a signal common to $\underline{X}$ and $\underline{Y}$ at lag L can be thought of as detecting a substring of $Z_1,\ldots,Z_N$ such that the probability of a success within the substring is higher than the probability of a success outside it.

The procedure proposed for the detection of the signal depends on two parameters p_0 and p_1, $p_0 \leq p_1$. p_0 is the probability of a success in the absence of a signal. p_1 is a lower bound for the probability of a success in the signal. p_0 and p_1 are specified by the investigator. It is sensible to take p_0 to lie within the conventional confidence intervals for the theoretical smear computed from proposition 3-1. p_1 should be close to 1.; the smaller the p_1, the larger the probability of a mismatch allowed by the investigator. It is desirable that the results of the procedure do not depend crucially on the choice of p_0 and p_1.

Suppose that $1 \leq i < j \leq N$ and let $L_{ij}(p_0,p_1)$ be the generalized log-likelihood ratio (GLLR) for the hypothesis testing problem

$$H_0: p = p_0 \quad \text{vs.} \quad H_A: p > p_1 \tag{5-2}$$

based on the substring

$$Z_i, Z_{i+1}\ldots, Z_j. \tag{5-3}$$

Let s_1 and s_0 be the number of successes and the number of mismatches and $\hat{p}=(s_1/s_0+s_1)$ be the fraction of matches in the substring (5-3). s_1 s_0 and $\hat{p}$ depend on i and j; the dependence is not indicated to avoid making subsequent expressions cumbersome. The generalized likelihood ratio (GLR) for the testing problem (5-2) based on the substring in (5-

3) is:

$$\frac{\underset{p>p_1}{\text{Sup}}\ p^{s_1}(1-p)^{s_0}}{p_0^{s_1}(1-p_0)^{s_0}}. \tag{5-4}$$

It can be easily verified that the function $s_1 \log p + s_0 \log(1-p)$ attains its maximum at $\hat{p}$, increases in $[0,\hat{p}]$ and decreases in $[\hat{p},1]$. Consequently,

$$\underset{p>p_1}{\text{Sup}}\ p^{s_1}(1-p)^{s_0} = \begin{cases} \hat{p}^{s_1}(1-\hat{p})^{s_0} & \text{if } \hat{p} \geq p_1 \\ p_1^{s_1}(1-p_1)^{s_0} & \text{if } \hat{p} \leq p_1 \end{cases} \tag{5-5}$$

and the GLLR for the hypothesis testing problem in (5-2) from the data in (5-3) is:

$$L_{ij}(p_0,p_1) = \begin{cases} s_1 \log \dfrac{\hat{p}}{p_0} + s_0 \log \dfrac{1-\hat{p}}{1-p_0} & \text{if } \hat{p} \geq p_1 \\ s_1 \log \dfrac{p_1}{p_0} + s_0 \log \dfrac{1-p_1}{1-p_0} & \text{if } \hat{p} \leq p_1 \end{cases} \tag{5-6}$$

For the specified p_0 and p_1, the proposed procedure finds the substring of (5-1) which maximizes the GLLR (5-6) over all the substrings on the diagonal (5-1). If the maximum GLLR exceeds a critical value which depends on N, p_0 and p_1 and will be determined by simulation later in this chapter, the procedure detects a signal common to words $\underline{X}$ and $\underline{Y}$ to show up as the substring maximizing the GLLR. Formally, if,

$$M(p_0,p_1) = \underset{1 \leq i < j \leq N}{\text{Max}} L_{ij}(p_0,p_1), \tag{5-7}$$

$$L_{IJ}(p_0,p_1) = M(p_0,p_1) \tag{5-8}$$

and $M(p_0,p_1)$ is greater than a critical value to be elaborated upon later, the proposed procedure detects the substring

$$Z_I,\ldots,Z_J, \tag{5-9}$$

to be a signal allowing for error with probability less than $1-p_1$, immersed in noise where the probability of a match equals p_0. We shall say that the signal is realized as the pair of substrings

$$X_I,\ldots, X_J$$

and

$$Y_{I+L},\ldots,Y_{J+L},$$

in the data.

The proposed procedure can be considered as a modified GLR for testing the hypothesis that $Z_1,\ldots,Z_N$ is a noisy string vs. the hypothesis that somewhere in the string there exists a signal of success probability higher than p_1. The relation between the two is examined in Appendix 1.

Remark that if for the substring in (5-9), $\hat{p}\geq p_2\geq p_1$, then

$$\sup_{p\geq p_1} p^{s_1}(1-p)^{s_0} = \sup_{p\geq p_2} p^{s_1}(1-p)^{s_0}$$

and therefore $L_{IJ}(p_0,p_1)=L_{IJ}(p_0,p_2)$; the same substring will maximize the GLLR for the choices (p_0,p_1) and (p_0,p_2).

When $p_1=p_0$, $L_{ij}(.,.)$ reduces to:

$$L_{ij}(p_0)=\begin{cases} s_1 \log \dfrac{\hat{p}}{p_0} + s_0 \log \dfrac{1-\hat{p}}{1-p_0} & \text{if } \hat{p} \geq p_0 \\ 0 & \text{if } \hat{p} \leq p_0 \end{cases}$$

which is well known to be the GLLR test statistic for the hypothesis

$$H_0: p = p_0 \quad \text{vs.} \quad H_A: p > p_0$$

The proposed procedure was applied to the diagonals at lags 5, 0

and -12 of the CC matrix of chorion proteins 292 and 18B. The diagonals were listed in table 4-3; at levels $\alpha_1 = \alpha_2 = .025$, their diagonal smear was significantly higher than the matrix smears for the chorion 292 and 18B proteins. p_0 was chosen at .10 and .17, the endpoints of the 99% confidence interval for the matrix smear given in table 4-2. For each p_0, p_1 was selected at p_0 and .70, .80, .90, .95.

Table 5-1 presents the substrings of proteins 292 and 18B that maximize the GLLR for the various choices of p_0 and p_1. In the discussion pertaining to tables 5-1, 5-2 and 5-3 the substrings of table 5-1 will be called signals; the critical values that the maximum GLLR will have to exceed for the substrings to be legitimately considered realizations of signals in the data will be elaborated upon later. The detection of the signal in the data depends on the choice of p_0 and p_1, but the same pair of substrings may maximize the GLLR for two different choices of (p_0,p_1). Next to each pair of similar substrings of table 5-1 is typed a 2 by 5 matrix of characters 0 and 1 is typed. The indices of the matrix elements correspond to the 2 by 5 choices for (p_0,p_1) and a matrix element is 1 if the substring listed maximizes the GLLR for the values of (p_0,p_1) specified by the indices of the matrix element.

Table 5-1. The substrings of chorion 292 and 18B proteins maximizing the GLLR of (5-6) for the 2 by 5 choices for (p_0,p_1). "1"s indicate the values of (p_0,p_1) for which the listed substrings maximize the GLLR.

LAG		
5	GGLGYEG	11111
	GGLGYEG	11111
0	The first 114 amino acids of both proteins	10000
		00000
0	MSTFAFLFLCIQACLVQNVFGVCRGGLGLKGLAAPACGCGGLGYEGLGY	01000
	MSTFAFLLLCAQACLIQSVYSYGCGCGCGGLGGYGGLGYGGLGYGGLGY	00000
0	MSTFAFLFLCIQACLVQNV	00100
	MSTFAFLLLCAQACLIQSV	11100
0	MSTFAFLFLCIQACL	00011
	MSTFAFLLLCAQACL	00011
-12	GSYGGEGIGNVAVAGELPVAGTTAVAGQVPIIGAVDFCGRANAGGCVSIGGRCTGCGCGCG	11110
	GEYGGTGIGNVAVAGELPVAGKTAVGGQVPIIGAVGFGGTAGAAGCVSIAGRCGGCGCGCG	11100
-12	YGGEGIGNVAVAGELPVAGTTAVAGQVPIIGAVDFCGRANAGGCVSIGGRCTGCGCGCG	00001
	YGGTGIGNVAVAGELPVAGKTAVGGQVPIIGAVGFGGTAGAAGCVSIAGRCGGCGCGCG	00011

Table 5-2 presents $\hat{p}$, the ratio of matches for the substrings of table 5-1 and table 5-3 lists the lengths of the diagonals and the values of M(.,.) for the substrings of table 5-1.

Table 5-2. Ratio of matches for the substrings of table 5-1.

			p_1			
LAG		p_0	.70	.80	.90	.95
5	p_0=.10	1.	1.	1.	1.	1.
	p_0=.17	1.	1.	1.	1.	1.
0	p_0=.10	.38	.53	.80	.87	.87
	p_0=.17	.80	.80	.80	.87	.87
-12	p_0=.10	.82	.82	.82	.82	.83
	p_0=.17	.82	.82	.82	.83	.83

Table 5-3. Values of M(.,.) attained for the 2 by 5 possible values of (p_0,p_1).

LAG	DIAGONAL LENGTH		p_1 .70	.80	.90	.95
		p_0				
5	116	p_0=.10	16.12	16.12	16.12	16.12	16.12
		p_0=.17	12.40	16.12	16.12	16.12	16.12
0	121	p_0=.10	30.95	25.33	25.18	24.17	23.49
		p_0=.17	17.55	17.55	17.54	17.43	16.75
-12	121	p_0=.10	87.50	87.50	87.50	85.69	81.41
		p_0=.17	61.86	61.86	61.86	60.50	56.22

Notice that for fixed p_0, as p_1 increases (i.e., as the procedure allows for a smaller probability of error in the signal) the substrings maximizing the GLLR are shorter and have a higher ratio of matches.

At lag 5 the heptapeptide GGLGYEG maximizes the GLLR of equation (5-6) for all selected values of (p_0,p_1). Depending on the values of p_0 and p_1, two different substrings on the diagonal at lag -12 maximize the GLLR. The signal detected for (p_0,p_1) = (.10, .95), (.17,.90) or (.17,.95) deletes from the longer signal - detected for the remaining seven values of (p_0,p_1) - its starting dipeptide which contains one mismatch. On the diagonal at lag 0 four different substrings maximize the GLLR for the ten choices of p_0 and p_1. A visual examination of the substring detected for p_0=.10 and p_1=.70 reveals that MSTFAFL*LC*QACL and GGLGY*GLGY are present on its right and left ends. (The occasional errors in the common string are denoted by an asterisk.) The substring on the left maximizes the GLLR when small probabilities of error (p_1=.90 or.95) are allowed. Noise intervenes between the two strings. When large probabilities of error are allowed for (p_0=.10 and p_1=.70) the matches on the right and left cover for the noise in the middle and the two signals together with the intervening noise maximize the GLLR. Similarly, a few

matches to the right of the string GGLGY*GLGY on the diagonal at lag 0 cause the substrings consisting of the first 114 amino acids of proteins 292 and 18B to maximize the GLLR for p=.10 vs. p>.10.

The BNC1 matrix for chorion proteins 292 and 18B was presented in figure 4-5; its visual examination was conducted prior to and independently of the application of the procedure proposed in the present chapter. Table 5-4 summarizes the results of the visual examination along lags 5, 0 and -12. When a prominent substring along the diagonal of the BNC1 matrix begins or ends with a "*", the realizations of the signal are taken to start or end one character to the left or right of "*".

Table 5-4. Strings shared in common by chorion 292 and 18B proteins recognized visually at the diagonals of table 4-3.

LAG			
5	GGLGYEGLG		
	GGLGYEGTG		
0	MSTFAFLFLCIQACLVQ	and	GGLGYEGLGY
	MSTFAFLLLCAQACLIQ	and	GGLGYGGLGY
-12	GSYGGEGIGNVAVAGELPVAGTTAVAGQVPIIGAVDFCGRANAGGCVSIGGRCTGCGCGCG		
	GEYGGTGIGNVAVAGELPVAGKTAVGGQVPIIGAVGFGGTAGAAGCVSIAGRCGGCGCGCG		

The visual examination of the BNC1 matrix of proteins 292 and 18B detects common substrings that are selected by the proposed procedure for the 2 by 5 choices for (p_0, p_1). The advantage of the proposed procedure is that it automates and quantifies the detection process.

The application of the proposed procedure may result in the two types of error that were referred to in chapter 4. Relating to the detection of a substring common to two words is the problem of the detection of a string of successes (up to a few occasional mismatches) in the word of (5-1). The latter will be called the one-dimensional problem to be contrasted from the former two-dimensional problem. The two types

of error are investigated for the one-dimensional problem first.

The asymptotic distribution of the GLLR $L_{ij}(p_0,p_1)$ of (5-6) for the hypothesis testing problem in (5-2) has been derived in [2]. The .95 quantile of the distribution of M(.,.) are estimated by simulation.

100 binary strings of length 50, 100, 200 and 300 were randomly generated with probabilities of success π = .10, .15 and .20. For each string M(.,.) was computed for p_0=.10, .15, .20 and p_1= p_0, .70, .80, .90, .95. To facilitate the presentation of the simulation results, the estimate of the .95 quantile of the distribution of $M(p_0,p_1)$ for noisy strings of length L generated with probability of success π is denoted by $Q_L(p_0,p_1|\pi)$. Tables 5-5 to 5-8 present the estimates $Q_{\cdot}(.,.|.)$ for each combination of string length L, probability of success in the binary strings π, and the nominal parameters p_0 and p_1. The distribution of M(.,.) is discrete. The .95 quantiles are estimated by the midpoint between the ninety-fifth and ninety-sixth largest observations for each combination of parameters. Next to $Q_{\cdot}(.,.|.)$, the largest observation from the 100 runs is listed in parenthesis in tables 5-5 to 5-8.

Table 5-5. Upper 5% points for M(.,.) estimated from 100 binary strings of length 50 generated with success probability π. The largest observations in the 100 runs are listed in parenthesis.

			p_1			
π		p_0	.70	.80	.90	.95
.10	p_0=.10	6.91(9.21)	6.91(9.21)	6.91(9.21)	6.91(9.21)	6.91(9.21)
	p_0=.15	5.69(7.59)	5.69(7.59)	5.69(7.59)	5.69(7.59)	5.69(7.59)
	p_0=.20	4.83(6.44)	4.83(6.44)	4.83(6.44)	4.83(6.44)	4.83(6.44)
.15	p_0=.10	7.82(12.1)	7.52(12.0)	6.97(11.5)	6.91(11.5)	6.91(11.5)
	p_0=.15	6.40(9.49)	5.69(9.49)	5.69(9.49)	5.69(9.49)	5.69(9.49)
	p_0=.20	4.83(8.05)	4.83(8.05)	4.83(8.05)	4.83(8.05)	4.83(8.05)
.20	p_0=.10	11.7(13.8)	9.23(13.8)	9.21(13.8)	9.00(13.8)	8.79(13.8)
	p_0=.15	7.59(11.4)	7.59(11.4)	7.26(11.4)	7.20(11.4)	7.00(11.4)
	p_0=.20	6.08(9.66)	6.00(9.66)	6.00(9.66)	5.94(9.66)	5.73(9.66)

Table 5-6. Upper 5% points for M(.,.) estimated from 100 binary strings of length 100 generated with success probability π. The largest observations in the 100 runs are listed in parenthesis.

			p_1			
π		p_0	.70	.80	.90	.95
.10	p_0=.10	7.22(9.21)	7.22(9.21)	7.15(9.21)	6.91(9.21)	6.91(9.21)
	p_0=.15	5.69(7.59)	5.69(7.59)	5.69(7.59)	5.69(7.59)	5.69(7.59)
	p_0=.20	4.83(6.44)	4.83(6.44)	4.83(6.44)	4.83(6.44)	4.83(6.44)
.15	p_0=.10	9.21(10.9)	9.21(10.1)	9.21(9.21)	9.21(9.21)	9.21(9.21)
	p_0=.15	7.59(7.59)	7.59(7.59)	7.59(7.59)	7.59(7.59)	7.59(7.59)
	p_0=.20	6.44(6.44)	6.44(6.44)	6.44(6.44)	6.44(6.44)	6.44(6.44)
.20	p_0=.10	13.1(17.2)	11.4(12.7)	10.0(11.5)	9.21(11.5)	9.21(11.5)
	p_0=.15	9.34(10.3)	7.86(9.49)	7.59(9.49)	7.59(9.49)	7.59(9.49)
	p_0=.20	6.56(8.05)	6.44(8.05)	6.44(8.05)	6.44(8.05)	6.44(8.05)

Table 5-7. Upper 5% points for M(.,.) estimated from 100 binary strings of length 200 generated with success probability π. The largest observations in the 100 runs are listed in parenthesis.

			p_1			
π		p_0	.70	.80	.90	.95
.10	p_0=.10	6.91(9.21)	6.91(9.21)	6.91(9.21)	6.91(9.21)	6.91(9.21)
	p_0=.15	5.69(7.59)	5.69(7.59)	5.69(7.59)	5.69(7.59)	5.69(7.59)
	p_0=.20	4.83(6.44)	4.83(6.44)	4.83(6.44)	4.83(6.44)	4.83(6.44)
.15	p_0=.10	10.5(15.4)	9.21(15.4)	9.06(15.4)	9.00(15.4)	8.79(15.1)
	p_0=.15	7.26(12.2)	7.26(12.2)	7.26(12.2)	7.20(12.2)	7.00(11.9)
	p_0=.20	6.00(9.96)	6.00(9.96)	6.00(9.96)	5.94(9.95)	5.73(9.69)
.20	p_0=.10	18.2(28.1)	11.5(16.1)	11.5(16.1)	11.0(16.1)	10.3(16.1)
	p_0=.15	9.49(13.3)	9.16(13.3)	9.15(13.3)	8.48(13.3)	7.91(13.3)
	p_0=.20	7.33(11.3)	7.31(11.3)	6.97(11.3)	6.69(11.3)	6.51(11.3)

Table 5-8. Upper 5% points for M(.,.) estimated from 100 binary strings of length 300 generated with success probability π. The largest observations in the 100 runs are listed in parenthesis.

π		p_1 = p_0	.70	.80	.90	.95
.10	p_0=.10	8.98(11.1)	8.24(11.1)	8.06(11.1)	8.06(11.1)	8.06(10.6)
	p_0=.15	6.64(8.76)	6.64(8.76)	6.64(8.76)	6.64(8.76)	6.64(8.76)
	p_0=.20	5.63(7.01)	5.63(7.01)	5.63(7.01)	5.63(6.95)	5.63(6.58)
.15	p_0=.10	12.2(16.0)	9.81(11.6)	9.40(11.5)	9.21(11.0)	9.21(10.6)
	p_0=.15	7.59(7.59)	7.59(7.59)	7.59(7.59)	7.59(7.59)	7.59(7.59)
	p_0=.20	6.44(7.01)	6.44(7.01)	6.44(7.01)	6.44(6.95)	6.44(6.58)
.20	p_0=.10	24.5(26.6)	13.7(15.3)	11.5(13.8)	11.5(13.8)	11.5(13.8)
	p_0=.15	10.3(13.3)	9.49(11.4)	9.49(11.4)	9.49(11.4)	9.49(11.4)
	p_0=.20	8.05(9.66)	8.05(9.66)	8.05(9.66)	8.05(9.66)	8.05(9.66)

Tables 5-5 to 5-8 indicate that when $p_0 \geq \pi$ the estimate $Q_L(p_0,p_1|\pi)$ is stable over choices of p_1 at all string length. It is expected that as total string length increases, the quantiles of the distribution of M(.,.) increase, the increase (expected) to be more noticeable for shorter strings. This holds in all the 144 comparisons of estimates of quantiles in tables 5-5 to 5-8 with thirteen exceptions. In particular,

$$Q_{100}(.10,p_1|.10) > Q_{200}(.10,p_1|.10) \quad \text{for } p_1=.10,.70,.80$$

$$Q_{100}(.15,p_1|.15) > Q_{200}(.15,p_1|.15) \quad \text{for } p_1=.15,.70,.80,.90,.95$$

and $$Q_{200}(.20,p_1|.15) > Q_{200}(.20,p_1|.15) \quad \text{for } p_1=.20,.70,.80,.90,.95.$$

An examination of the simulation data of length 200 indicates that for all the above values of π, p_1 and p_2 either the ninety-fifth or the ninety-sixth largest simulated observations equal $Q_{100}(p_0,p_1|\pi)$. The discrepancy is minor and does not deserve further attention.

Note that when $\pi > p_0$, $Q_{.}(p_0,p_1|\pi) > Q_{.}(p_0,p_1|p_0)$. As a result, when $Q_{.}(p_0,p_1|p_0)$ is used as a critical threshold and noisy data are produced with success probability $\pi > p_0$, the probability of a type I error (false alarm) becomes considerably higher. For example, for noisy strings of length 50 generated with success probabilities .15 and .20 the

probabilities $Pr\{M(.10,.95) \geq Q_{50}(.10,.95|.10\}$ are estimated to be .15 and .30. For noisy strings of length 200 and success probabilities .15 and .20, $Pr\{M(.10,.95) \geq Q_{200}(.10,.95|.10\}$ is estimated to be .43 and .67 respectively.

The complete statistical assessment of the procedure requires the investigation of the probabilities of "false alarm" in conjunction with that of no detection of a signal present in the data. To this purpose, 50 strings of length L= 50, 100, 200 and 300 were generated. The strings consisted of signals of lengths S= 5, 7, 9, 11, 13 and 15 of probability of success σ= .90 implanted into noise of success probability π= .15. In the remainder of the chapter signals and noise will be understood to be Bernoulli variables with success probabilities σ= .90 and π= .15 respectively. Signals were implanted at one tenth and half of the noisy string length. Values used as critical thresholds will be explained further on. If for some run M(.,.) exceeds the critical threshold value, the substring for which $M(p_0,p_1)$ is attained, is detected by the procedure.

Use of (α,β) curves is made to present the performance of the proposed procedure when applied with parameters p_0=.10,.15,.20 and p_1=.80,.90,.95 on the simulation data. The (α,β) curves for the detection of signals of length S implanted in noisy strings of length L-S are curves passing through the points (α_i,β_i). α_i and β_i correspond to the choice of several critical thresholds C_i. Criteria C_i were chosen to be the midpoints between the values attained by the maximum GLLR $M(p_0,p_1)$ for the 100 noisy strings and the 50 strings where signals were implanted. α_i is the estimated probability of a "false alarm" when the procedure with parameters p_0 and p_1 and critical threshold C_i is applied

to noisy data of length L. β_i is the estimated probability of no detection when the same procedure (with parameters p_0, p_1 and C_i) is applied to noisy strings within which signals of length S have been implanted, as explained. α_i depends on π, the test parameters C_i, p_0, and p_1 and the total signal length L. In addition to these parameters, β_i depends on σ (σ= .90 in this study), signal length S and the position in which the signal is implanted within the overall string. This dependence of α_i and β_i is not explicitly denoted to avoid making expressions cumbersome.

Figure 5-1a presents the nine (α,β) curves corresponding to the choices p_0=.10,.15,.20 and p_1=.80,.90,.95 for the detection of a signal of length 5 implanted at the first tenth of the noisy string of 45 characters. Figures 5-1b and 5-1c plot the same curves for signals of length 7 and 9 implanted at the first tenth of noise, the overall strings being 50 characters long. Figure 5-1d plots all 27 curves in the same frame. Figures 5-2a to 5-2d present the same plots for signals of length 5, 7 and 9 implanted in noisy strings at the first tenth of noise, overall strings being 300 characters long. Figures 5-3 and 5-4 plot the corresponding curves for signals of length 5,7,9 implanted at the middle of noise, overall strings being 50 and 300 characters long.

The proposed procedure is rather powerless in detecting signals of length 5 implanted in noisy strings of 295 characters. When the signal is implanted at the first tenth of the total string length, for all nine values of (p_0,p_1) p_0=.10,.15,.20 and p_1=.80,.90,.95 there is no critical threshold value C_i for which the two estimated probabilities of error α_i and β_i are both less than 15%. This is illustrated in figure 5-2a; lying on the unit square, the (α,β) curves do not cross the square

[0.,.15]x[0.,.15]. Figure 5-4a illustrates that when a signal of 5 characters is implanted at the middle of the overall string of 300 characters, for no critical thresholds are the estimates of the probabilities of two kinds of error less than 20% because it is very likely that in a noisy string of 300 characters and success probability there will exist a string of no less than four successes. Neither is the procedure particularly powerful in detecting a signal of five characters implanted within a noisy string 45 characters. When the signal is implanted at the first tenth of the overall string, there are criteria C_i for which both α_i and β_i are both less than 15%, but not less than 10%. When the signal is implanted at the middle of the overall string length, for C=6.98, p_0=.10 and p_1=.80, α=.05 and β=.08.

The curves in figures 5-1c, 5-2c, 5-3c and 5-4c indicate that the proposed procedure is quite powerful in detecting signals of length 9 implanted in noisy stings of length 45 and 295. Since the scales in which the (α,β) curves are drawn do not allow the estimates of the probabilities of the two kinds of errors to be read, test parameters (critical threshold C, p_0 and p_1) for which estimated probabilities for the two kinds of errors are small, are presented in tables 5-9 to 5-12.

Table 5-9. Critical values and estimates of the probabilities of the two kinds of errors when detecting a signal of length 9 implanted at the first tenth of a noisy string of 41 characters by $M(p_0,p_1)$. π=.15,σ=.90

	p_1 = .80			.90			.95		
	C	α	β	C	α	β	C	α	β
p_0=.10	6.98	.05	.0	7.85	.04	.0	7.32	.04	.0
	7.89	.04	.0	9.00	.03	.02	8.05	.04	.02
	9.19	.03	.0	10.1	.03	.04	8.79	.03	.02
	9.89	.03	.02	11.3	.02	.04	9.59	.03	.04
	10.9	.03	.04	12.6	.0	.06	10.7	.02	.04
	12.6	.00	.06	14.6	.0	.08	12.6	.0	.06
p_0=.15	6.31	.04	.0	6.08	.04	.0	6.05	.04	.02
	7.05	.03	.0	6.65	.04	.02	6.99	.03	.02
	7.37	.03	.02	7.20	.03	.04	8.54	.02	.04
	8.20	.03	.04	7.93	.03	.04	10.4	.0	.06
	9.16	.02	.04	8.81	.02	.04			
	10.4	.0	.06	10.4	.0	.06			
p_0=.20	5.19	.04	.0	4.85	.04	.0	4.93	.04	.02
	5.56	.04	.02	5.15	.04	.02	5.73	.03	.02
	6.01	.03	.02	5.94	.03	.02	7.24	.02	.04
	6.69	.03	.04	7.24	.02	.04	8.85	.0	.06
	7.49	.02	.04	8.85	.0	.08			
	8.85	.0	.0						

Table 5-10. Critical thresholds and estimates of the probabilities of the two kinds of errors when detecting a signal of length 9 implanted at the first tenth of a noisy string of 291 characters by $M(p_0,p_1)$. π=.15, σ=.90

	p_1 = .80			.90			.95		
	C	α	β	C	α	β	C	α	β
p_0=.10	9.27	.06	.02	10.1	.02	.04	9.59	.02	.04
	9.73	.03	.02	11.2	.0	.14	10.3	.01	.14
	10.5	.02	.04				11.5	.0	.18
	11.3	.01	.04						
p_0=.15	7.04	.17	.02	7.92	.02	.04	7.42	.13	.06
	7.26	.15	.02	8.43	.01	.14	7.91	.01	.14
	7.48	.14	.04	9.05	.0	.14	8.86	.0	.14
	8.13	.02	.04						
	8.74	.01	.04						
p_0=.20	6.00	.14	.04	5.90	.14	.04	5.19	.13	.06
	6.68	.02	.04	6.40	.13	.06	6.51	.01	.14
	6.97	.01	.14	6.69	.01	.14			
	7.73	.0	.14						

AD-A153 605 TOWARDS A STATISTICAL ANALYSIS OF GENETIC SEQUENCES 2/2
DATA WITH PARTICULAR..(U) MASSACHUSETTS INST OF TECH
CAMBRIDGE STATISTICS CENTER S P ARSENIS MAR 85
UNCLASSIFIED TR-36-ONR N00014-74-C-0555 F/G 6/3 NL

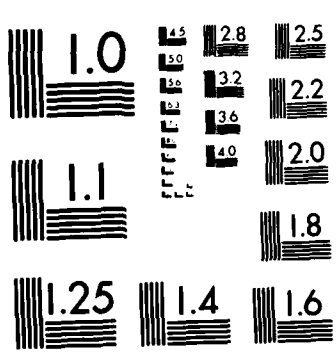


Table 5-11. Critical thresholds and estimates of the probabilities of the two kinds of errors when detecting a signal of length 9 implanted at the middle of a noisy string of 41 characters by $M(p_0,p_1)$. π=.15, σ=.90

	p_1								
	.80			.90			.95		
	C	α	β	C	α	β	C	α	β
p_0=.10	7.50	.04	.0	9.00	.03	.02	8.79	.03	.02
	8.43	.04	.02	10.1	.03	.04	9.59	.03	.04
	9.76	.03	.02	11.3	.02	.06	10.3	.02	.04
	10.9	.03	.04	12.6	.0	.10	11.1	.02	.06
	11.3	.03	.06	13.5	.0	.14	12.2	.0	.10
p_0=.15	6.32	.04	.02	7.20	.03	.02	6.99	.03	.02
	7.27	.03	.02	7.93	.03	.04	7.91	.02	.04
	8.14	.03	.04	8.84	.02	.04	8.86	.02	.06
	8.75	.03	.06	9.05	.02	.06	9.80	.0	.10
	9.16	.02	.06	9.94	.0	.10			
p_0=.20	5.20	.04	.02	5.14	.04	.02	4.93	.04	.02
	6.01	.03	.02	5.94	.03	.02	5.73	.03	.02
	6.69	.03	.04	6.69	.02	.04	6.51	.02	.04
	6.97	.02	.04	7.50	.02	.06	7.31	.02	.06
	7.53	.02	.06						

Table 5-12. Critical thresholds and estimates of the probabilities of the two kinds of errors when detecting a signal of length 9 implanted at the middle of a noisy string of 291 characters by $M(p_0,p_1)$. π=.15, σ=.90

	p_1								
	.80			.90			.95		
	C	α	β	C	α	β	C	α	β
p_0=.10	9.27	.06	.02	10.1	.02	.0	9.59	.02	.0
	9.40	.05	.0	11.2	.0	.02	10.3	.01	.0
	9.73	.03	.0				11.0	.0	.02
	10.5	.02	.0						
	11.3	.01	.0						
p_0=.15	7.48	.14	.0	7.20	.14	.0	7.42	.13	.0
	8.13	.02	.0	7.92	.02	.0	7.91	.01	.0
	8.74	.01	.0	8.43	.01	.0	8.86	.0	.04
p_0=.20	6.68	.02	.0	6.40	.13	.0	6.51	.01	.0
	6.97	.01	.0	6.69	.01	.0	7.31	.0	.04
	7.53	.0	.02	7.12	.0	.02			

Since for all nine choices of (p_0, p_1), there are thresholds for which the probabilities of the two types of error are both less than .05 the proposed procedure is illustrated to be quite powerful and robust in detecting a signal of length 9 implanted in the middle of noisy strings as long as 291 characters. The procedure is weaker when the signal is implanted at the first tenth of the noisy string.

When a signal of 7 characters is implanted at the first tenth of a noisy string of 293 characters, for no values of test parameters C, p_0 and p_1, are the probabilities of both types of error less than 10%. When the signal is implanted in the middle of the noisy string (of 293 characters), it is only for p_0=.10 and p_1=.80 that both probabilities can become less than 10%. In particular for

C=9.40 α=.05 and β=.02

and C=9.73 a=.03 and β=.08.

The procedure is more powerful in detecting a signal of length 7 within noise 43 characters long. Test parameters for which the two types of error are less than 10% are given in table 5-13. Each cell of table 5-13 considered as a three-way table comprises of two triplets for C, α and β; the top for signals implanted at the first tenth of the noisy string and the bottom for signals implanted at the middle.

Table 5-13. Critical thresholds and estimates of the probabilities of the two kinds of errors when detecting a signal of length 7 implanted at one tenth (above) and the middle (below) of a noisy string of 43 characters by $M(p_0,p_1)$. π=.15, σ=.90.

		p_1 = .80			p_1 = .90			p_1 = .95	
	C	α	β	C	α	β	C	α	β
p_0=.10	6.98	.05	.0	7.85	.04	.04	7.64	.04	.04
	6.98	.05	.0	7.85	.04	.06	7.64	.04	.06
	7.22	.04	.0	9.00	.03	.06	8.79	.03	.06
	7.22	.04	.0	9.00	.03	.08	9.18	.03	.08
	8.15	.04	.04						
	7.68	.04	.04						
	9.06	.03	.04						
	8.43	.04	.06						
	9.34	.03	.08						
	9.19	.03	.06						
p_0=.15	6.31	.04	.04	6.26	.04	.04	6.05	.04	.04
	5.69	.04	.04	6.26	.04	.06	6.05	.04	.06
	7.05	.03	.04	7.20	.03	.06	6.99	.03	.06
	6.32	.04	.06	7.54	.03	.08	6.82	.03	.08
	7.37	.03	.06						
	7.05	.03	.06						
p_0=.20	5.20	.04	.04	5.14	.04	.04	4.93	.04	.04
	5.20	.04	.06	5.14	.04	.06	4.93	.04	.06
	6.01	.03	.06	5.94	.03	.06	5.73	.03	.06
	6.25	.03	.08	5.90	.03	.08	5.91	.03	.08

The two errors considered thus far were "false alarms" and no detection when a signal is present. It is possible however, that the procedure detects a signal but detection is not accurate. Detection is perfectly accurate when the substring maximizing the GLLR is identical to the implanted signal. However, given that errors are allowed within signals, perfectly accurate detection is overly restrictive; in analyzing the simulation data for accurate detection, allowance has to be made for moderate deviations between the two substrings. These deviations are measured in an ad hoc fashion by the sum of the distances between the beginning and endpoints of the two substrings. Formally, if the implanted

signal within $Z_1,Z_2,\ldots,Z_N$ is $Z_A,Z_{A+1},\ldots,Z_B$ and the substring maximizing $M(p_0,p_1)$is $Z_I,Z_{I+1},\ldots,Z_J$, the deviation between the two substrings is taken to be $D=|I-A|+|J-B|$. The detection of the implanted signal is considered accurate if the sum is not larger than the smallest integer larger than half the length of the implanted signal. In particular, signals of length 5,7,9,11,13 and 15 are considered to be accurately detected if the sum is not larger than 3,4,5,6,7 or 8. Since the performance of the proposed procedure in detecting a signal of length 5 is not satisfactory, its performance in detecting accurately will be examined only for signals of length 7,9 and 11.

Figures 5-5a, 5-5b and 5-5c plot the (α,β) curves for accurate detection of signals of length 7, 9 and 11 implanted at the first tenth of noisy strings, the overall string length being 50. Nine curves are plotted on each frame, corresponding to (p_0,p_1) for the choices p_0=.10,.15,.20 and p_1=.80,.90,.95. Figure 5-5d superimposes all 27 curves on the same frame. Figures 5-6 plot the same curves for signals of length 7,9 and 11 implanted in noise, the overall string length being 300.

On some of the plots on figures 5-5 and 5-6, the probability of accurate detection cannot be made larger than 98%, i.e. β cannot be made less than 2%, no matter how large the α, i.e. even for very small critical thresholds. This is so because a relatively large number of mismatches in the implanted signal may cause a substring of the noisy string to maximize the GLLR in the overall string. Figure 5-7 lists the substrings maximizing the GLLR and the maximum GLLR $M(p_0,p_1)$ attained for the nine choices of (p_0,p_1) when the procedure is applied to detect a signal implanted at sites 5 to 11 and the overall string length is 50 characters. With one exception marked on the figure, in all 50 runs the

substrings maximizing the GLLR are close to the implanted signal.

Since the scales on which figures 5-5 and 5-6 are drawn do not allow the probabilities of "false alarms" and no detection or non-accurate detection to be read off, criteria for which the estimated probabilities are small are listed in tables. Tables 5-14 and 5-15 list criteria for which the estimated probabilities of the two kinds of errors are small when the procedure is applied to detect accurately signals of length 7 and 9 implanted at the first tenth of noisy strings of length 43 and 41.

Table 5-14. Critical thresholds and estimates of the probabilities of the two kinds of errors when detecting accurately a signal of length 7 implanted at the first tenth of a noisy string of 43 characters by $M(p_0, p_1)$.

	p_1 = .80			.90			.95		
	C	α	β	C	α	β	C	α	β
p_0=.10	6.98	.05	.08	7.85	.04	.10	7.64	.04	.08
	7.22	.04	.08	9.00	.03	.12	8.79	.03	.16
p_0=.15	7.05	.03	.10	6.26	.04	.08	6.05	.04	.06
	7.37	.03	.12	7.20	.03	.10	6.99	.03	.08
p_0=.20	5.20	.04	.06	5.14	.04	.06	4.93	.04	.06
	6.01	.03	.08	5.94	.03	.08	5.73	.03	.08

Table 5-15. Critical thresholds and estimates of the probabilities of the two kinds of errors when detecting accurately a signal of length 9 implanted at the first tenth of a noisy string of 41 characters by $M(p_0, p_1)$.

	p_1 = .80			.90			.95		
	C	α	β	C	α	β	C	α	β
p_0=.10	7.98	.04	.02	10.1	.03	.04	8.79	.03	.02
	9.19	.03	.02	11.3	.02	.04	10.7	.02	.04
p_0=.15	7.37	.03	.02	7.20	.03	.02	6.99	.03	.02
	9.16	.02	.04	8.83	.02	.04	8.54	.02	.04
p_0=.20	6.01	.03	.02	5.94	.03	.02	5.73	.03	.02
	7.49	.02	.04	7.24	.02	.04	7.24	.02	.04

The procedure is quite powerful and robust; accurate detection is accomplished with errors of small probability (less than 5%) when the procedure parameters are selected to be p_0=.10,.15,.20 and p_1=.80,.90,.95 while the probabilities of success in noise and signal are .15 and .90. Since the performance of the proposed procedure in detecting a signal of length 7 implanted within a noisy string of length 293 was poor, the operating characteristics of the procedure are given for the accurate detection of signals of 9 and 11 characters only.

Table 5-16 Critical thresholds and estimates of the probabilities of the two kinds of errors when detecting accurately a signal of length 9 implanted at the first tenth of a noisy string of 291 characters by $M(p_0,p_1)$.

	p_1 = .80			.90			.95		
	C	α	β	C	α	β	C	α	β
p_0=.10	9.74	.03	.06	10.1	.02	.04	9.59	.02	.04
	10.6	.02	.08						
p_0=.15	8.75	.01	.04	7.93	.02	.04	7.42	.13	.06
							7.91	.01	.14
p_0=.20	6.69	.02	.04	6.40	.13	.06	5.90	.13	.06
				6.69	.01	.14	6.51	.01	.14

Table 5-17. Critical thresholds and estimates of the probabilities of the two kinds of errors when detecting accurately a signal of length 11 implanted at the first tenth of a noisy string of 289 characters by $M(p_0,p_1)$.

	p_1 = .80			.90			.95		
	C	α	β	C	α	β	C	α	β
p_0=.10	9.74	.03	.04	10.1	.02	.0	9.27	.02	.02
	10.6	.02	.06	12.1	.0	.02			
p_0=.15	8.75	.01	.02	8.44	.01	.04	7.91	.01	.06
p_0=.20	6.97	.01	.02	6.69	.01	.06	6.51	.01	.06

The probabilities of the two kinds of errors listed in tables 5-16 and 5-

10 for accurate detection and detection are not substantially different. The procedure is quite powerful in accurately detecting a signal of length 9 within a string of overall length 300 and, as expected, more powerful and remarkably robust in accurately detecting a signal of length 11.

The two-dimensional problem to detect a signal within two words $\underline{X}$ and $\underline{Y}$ is now briefly addressed. In examining character matrices visually, the investigator scans each diagonal for substrings with a few occasional mismatches. In an analogous fashion, the procedure for the two dimensional problem transforms blank and nonblank characters to 0's and 1's and computes the maximum GLLR along each diagonal for selected values of p_0 and p_1. Let the maximum GLLR along the matrix diagonal of lag k be denoted by $M(p_0,p_1,k)$. If $M(p_0,p_1,k)$ is larger than some critical value, the substrings of $\underline{X}$ and $\underline{Y}$ along the diagonal at lag k for which $M(p_0,p_1,k)$ is attained are considered to be realizations of a common signal in the data.

Except for the nominal parameters p_0 and p_1, the critical threshold should depend on the amino acid counts in $\underline{X}$ and $\underline{Y}$; it is chosen to be the estimated .95 quantile of the distribution of $\text{Max}\{M(p_0,p_1,k)\}$ for random permutations of the words.

The proposed procedure has been applied to chorion proteins 292 and 18B for p_0=.20 and p_1=.90. Figure 5-8 plots $M(p_0,p_1,k)$ at each matrix diagonal. For 100 permutations of protein 18B the 29 largest values of $\text{Max}\{M(p_0,p_1,k)\}$ are: 11.3, 9.66(3), 8.45, 8.05(24). (Numbers in parenthesis denote ties.) Hence, with 8.2 as a critical value the procedure detects nine signals in (nine) matrix diagonals. The ralizations of the signals in the data are listed in decreasing order in

$M(p_0,p_1,..)$ in table 5-18.

Table 5-18. Realizations of signals detected in proteins 292 and 18B.

LAG		$M(p_0,p_1)$
-12	YGGEGIGNVAVAGELPVAGTTAVAGQVPIIGAVDFCGRANAGGCVSIGGRCTGCGCGCG YGGTGIGNVAVAGELPVAGKTAVGGQVPIIGAVGFGGTAGAAGCVSIAGRCGGCGCGCG	52.9
0	MSTFAFLFLCIQACL MSTFAFLLLCAQACL	15.4
-5	GGLGYEGLGYGALGY GGLGYGGLGYGGLGY	15.4
-10	GYEGLGYGALGYDGLGY GYGGLGYGGLGYGGLGY	14.8
-15	GYGALGYDGLGYG GYGGLGYGGLGYG	12.4
5	GGLGYEG GGLGYEG	11.3
-11	CGCGGLG CGCGGLG	11.3
-100	GCGCGCG GCGCGCG	11.3
-20	GYDGLGYG GYGGLGYG	8.4

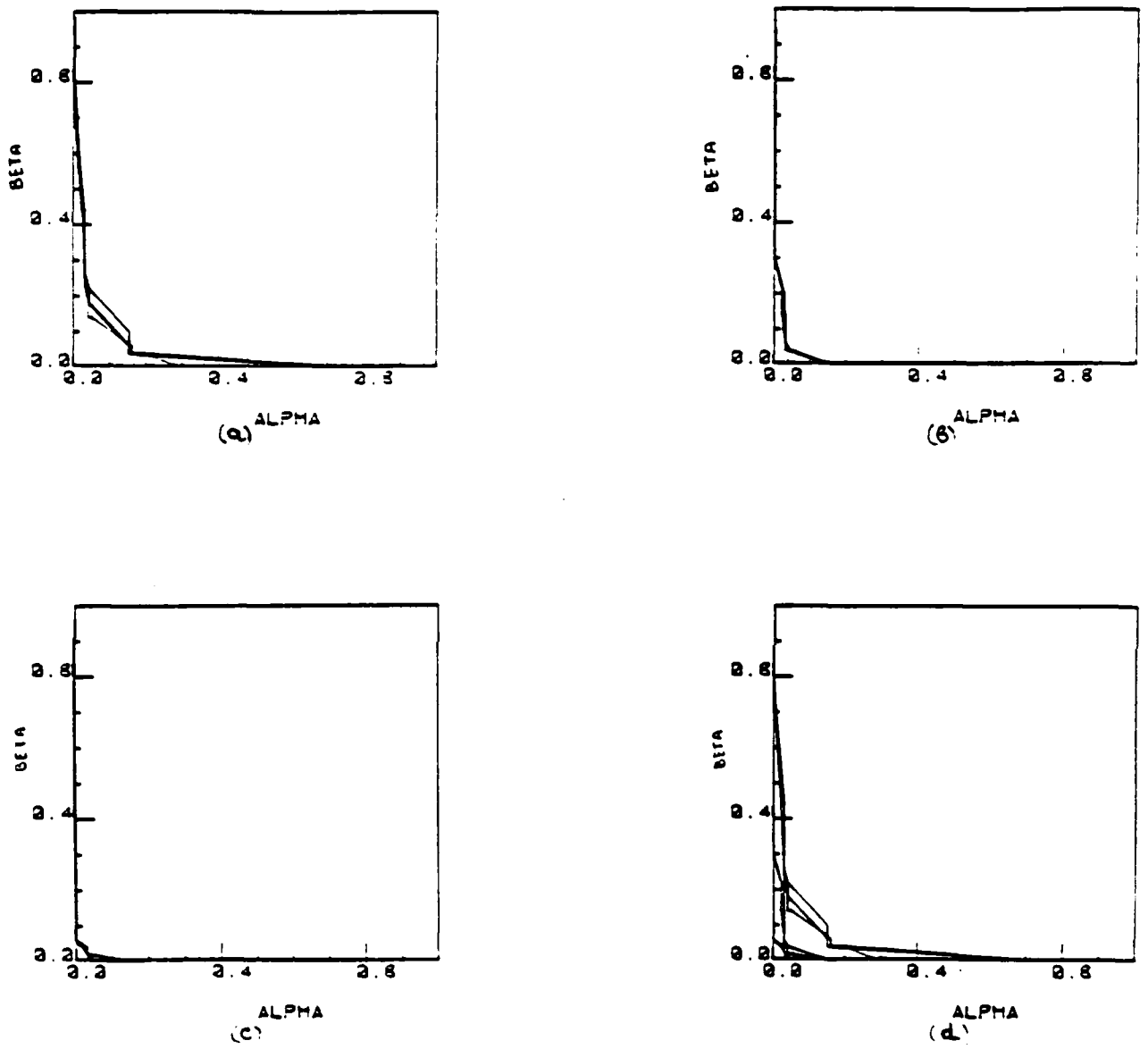


Figure 5-1.(α,β) curves for detection by $M(p_0,p_1)$ of a signal implanted within noise at first tenth of noisy string . Overall string length L=50. p_0=.10,.15,.20, and p_1=.80,.90,.95. π=.10, σ=.90. (a) Signal 5 characters long. (b) Signal 7 characters long. (c) Signal 9 characters long. (d) superimposes plots of a,b and c on one frame.

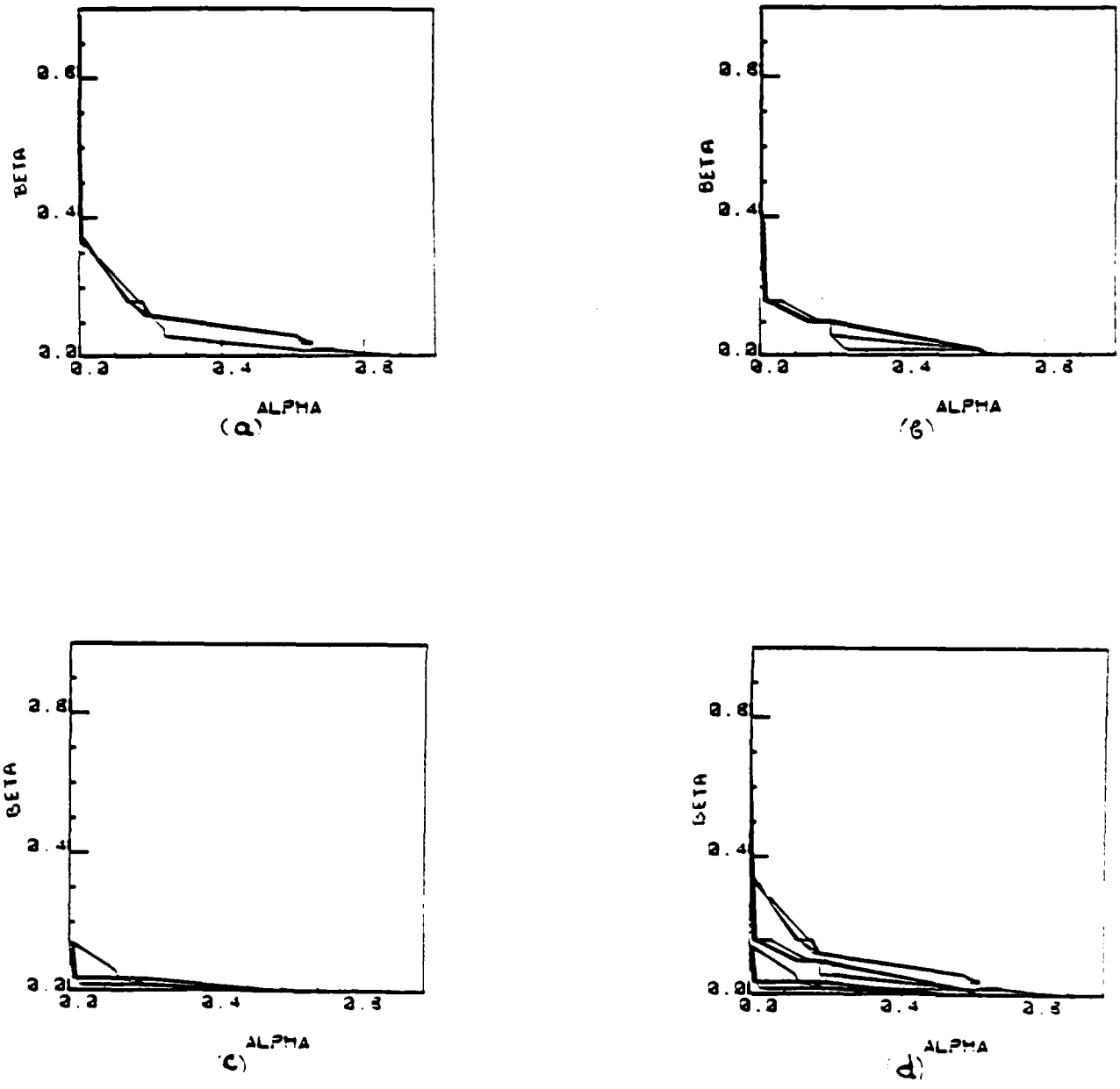


Figure 5-2. (α,β) curves for detection by $M(p_0,p_1)$ of a signal implanted within noise at first tenth of noisy string. Overall string length L=300. p_0=.10,.15,.20, and p_1=.80,.90,.95. π=.10, σ=.90. (a) Signal 5 characters long. (b) Signal 7 characters long. (c) Signal 9 characters long. (d) superimposes plots of a,b and c on one frame.

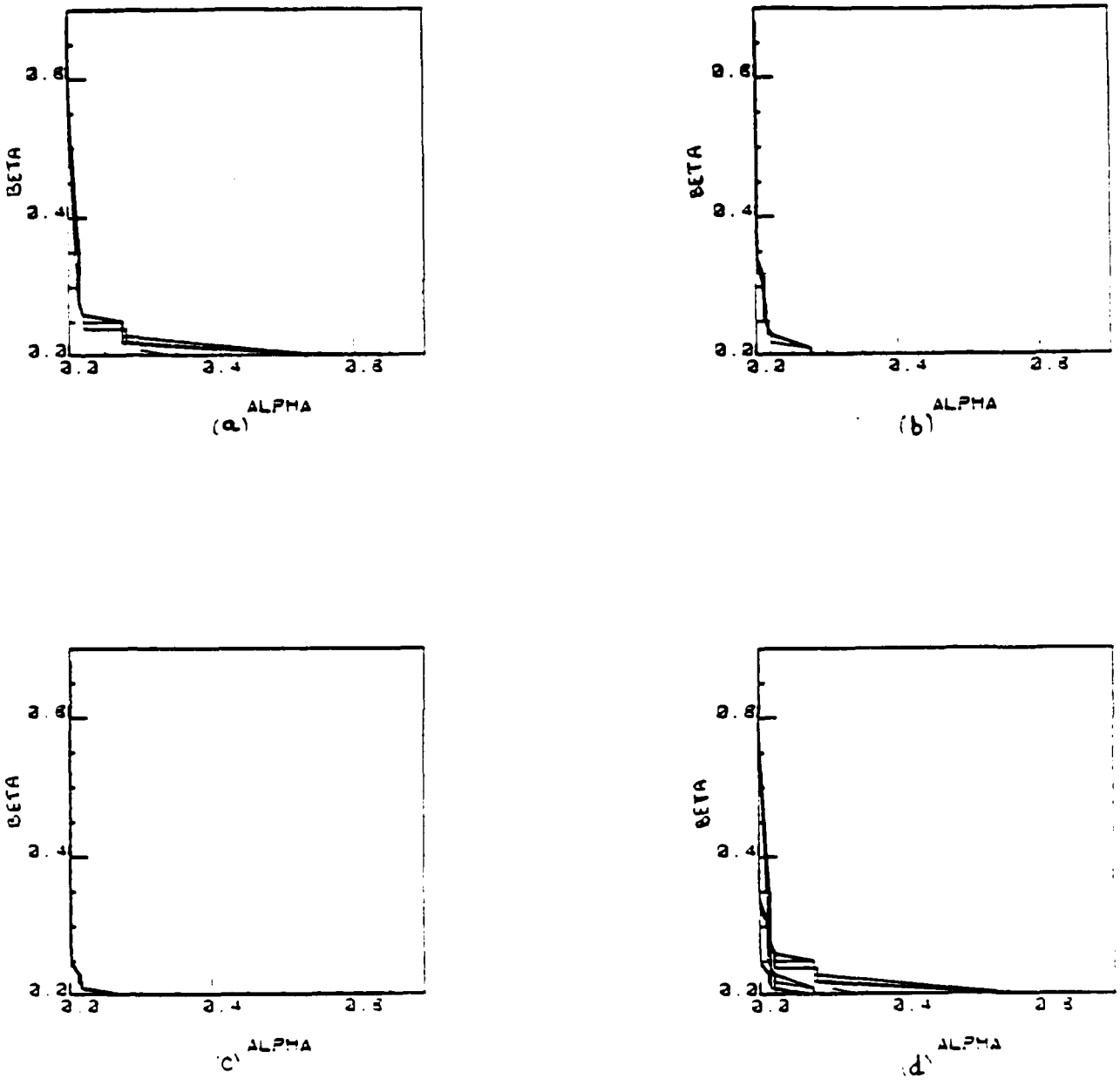


Figure 5-3. (α,β) curves for detection by $M(p_0,p_1)$ of a signal implanted within noise at the middle of noisy string . Overall string length L=50. p_0=.10,.15,.20, and p_1=.80,.90,.95. π=.10, σ=.90. (a) Signal 5 characters long. (b) Signal 7 characters long. (c) Signal 9 characters long. (d) superimposes plots of a,b and c on one frame.

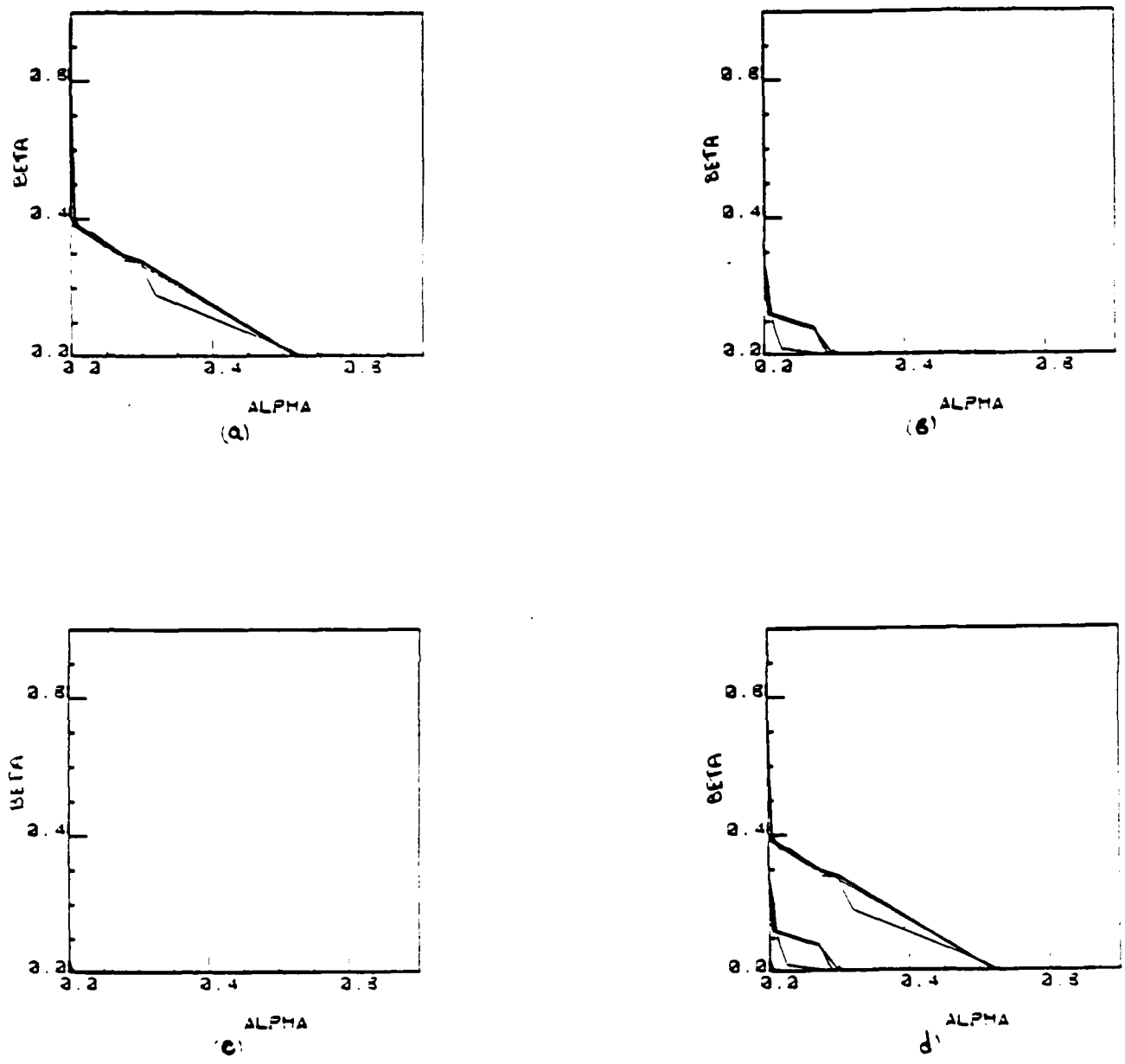


Figure 5-4. (α,β) curves for detection by $M(p_0,p_1)$ of a signal implanted within noise at the middle of noisy string. Overall string length L=300. p_0=.10,.15,.20, and p_1=.80,.90,.95. π=.10, σ=.90. (a) Signal 5 characters long. (b) Signal 7 characters long. (c) Signal 9 characters long. (d) superimposes plots of a,b and c on one frame.

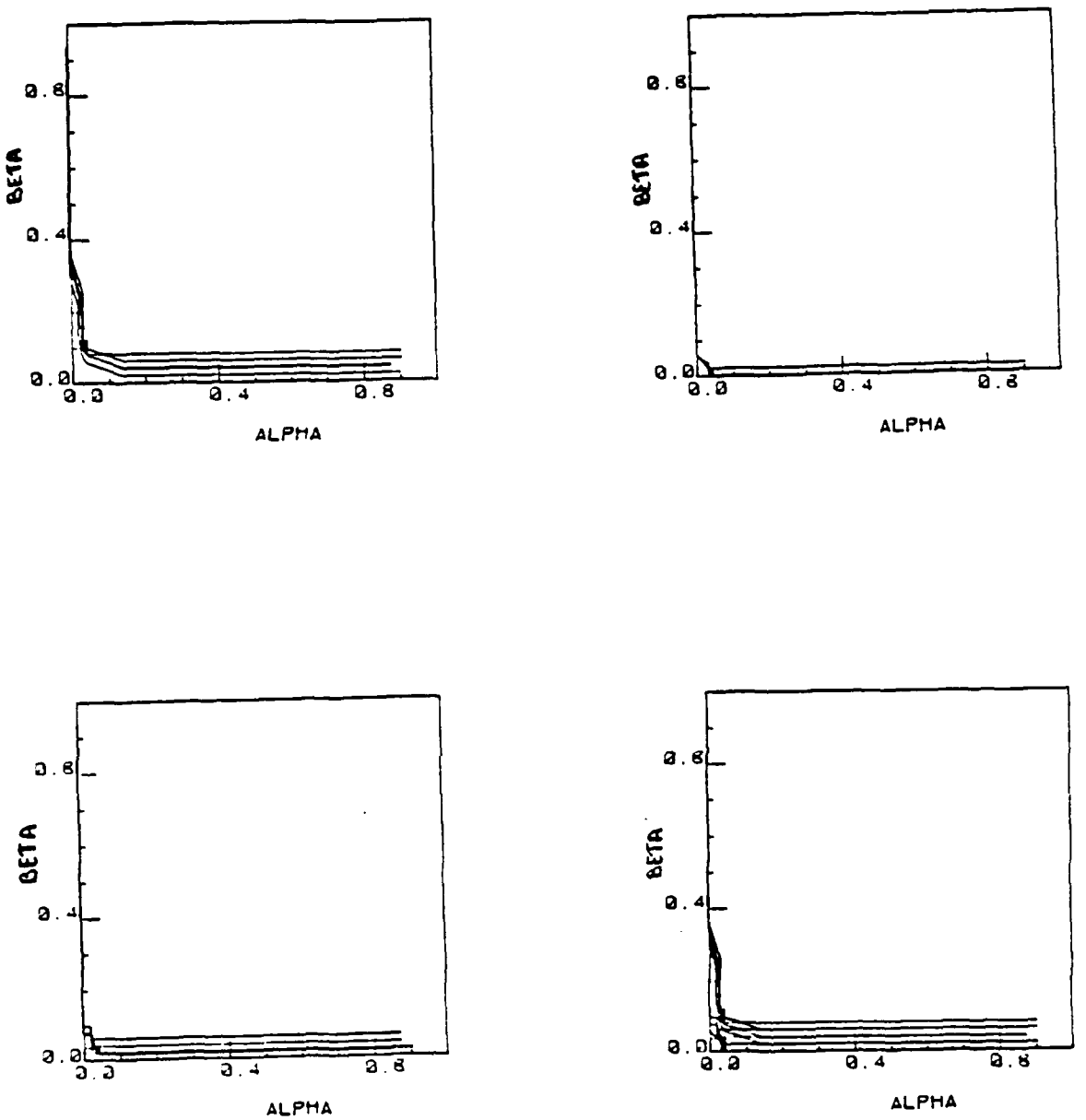


Figure 5-5. (α,β) curves for accurate detection by $M(p_0,p_1)$ of a signal implanted within noise at first tenth of noisily string. Overall string length L=50. p_0=.10,.15,.20, and p_1=.80,.90,.95. π=.10, σ=.90. (a) Signal 7 characters long. (b) Signal 9 characters long. (c) Signal 11 characters long. (d) superimposes plots of a,b and c on one frame.

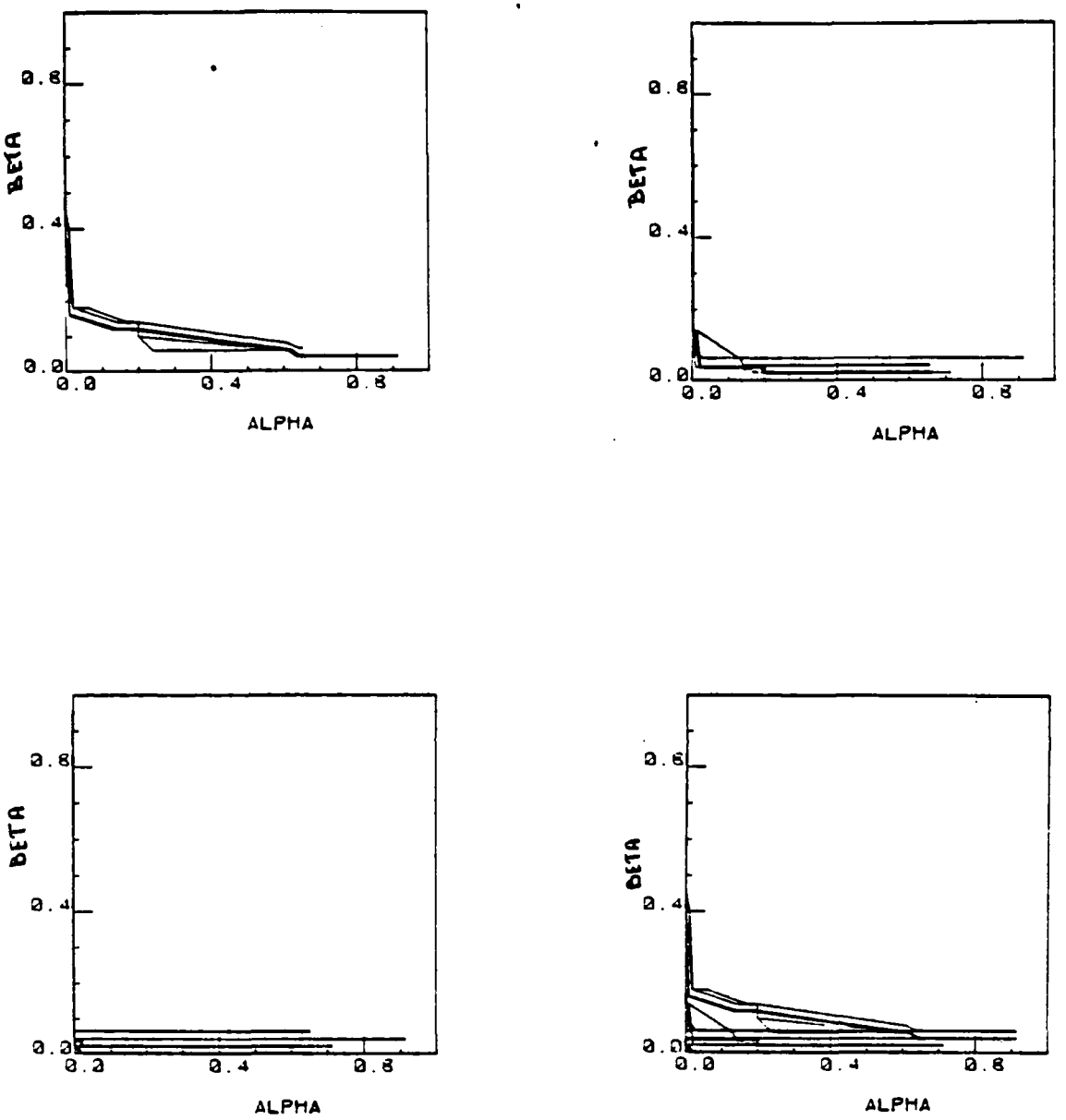


Figure 5-6. (α,β) curves for accurate detection by $M(p_0,p_1)$ of a signal implanted within noise at first tenth of noisy string. Overall string length L=300. p_0=.10,.15,.20, and p_1=.80,.90,.95. π=.10, σ=.90. (a) Signal 7 characters long. (b) Signal 9 characters long. (c) Signal 11 characters long. (d) superimposes plots of a,b and c on one frame.

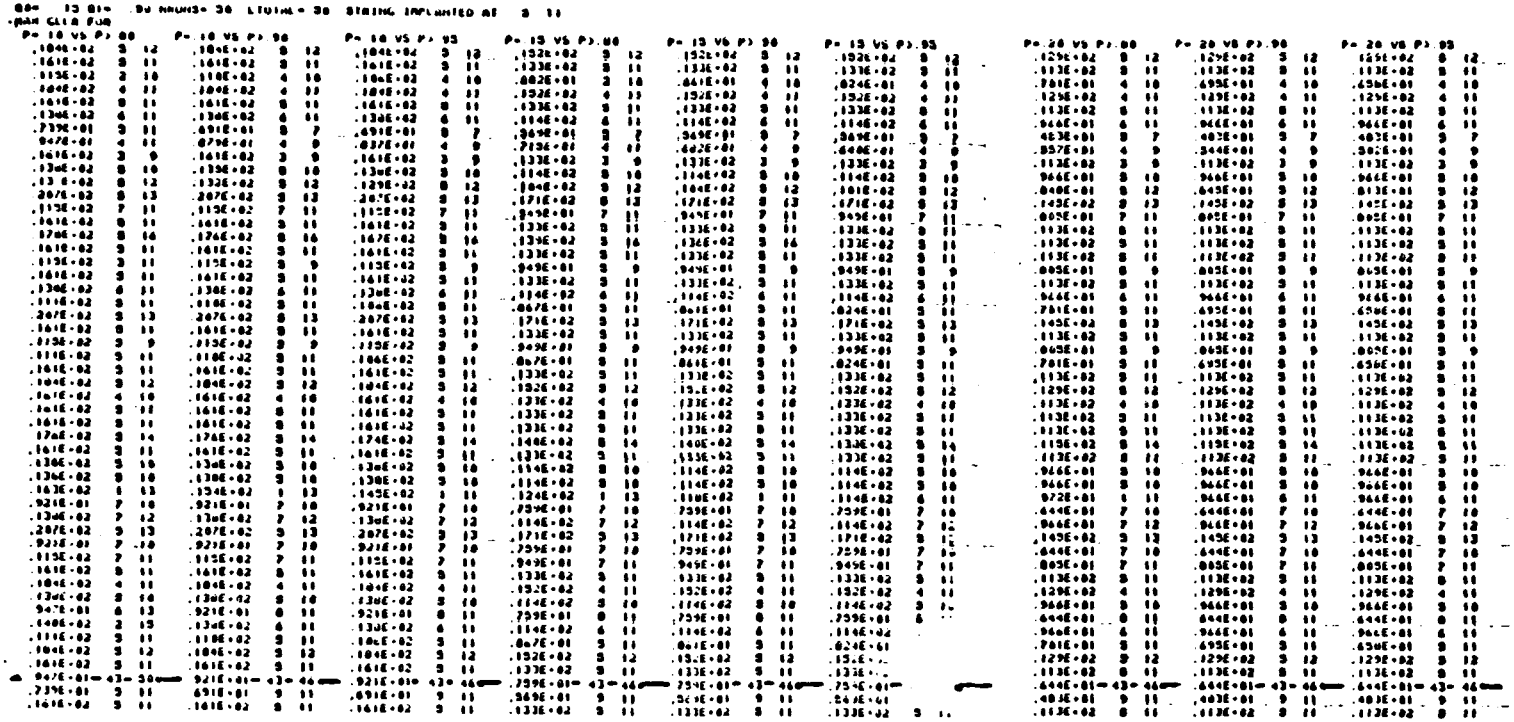


Figure 5-7. Beginning and endpoints of substrings maximizing the GLLR of equation (5-6) and values of the maximum GLLR M(.,.) for all nine values of (p_0, p_1), p_0=.10,.15,.20 and p_1=.80,.90,.95, to detect a signal of length 7 implanted at the first tenth of a noisily substring of 43 characters.

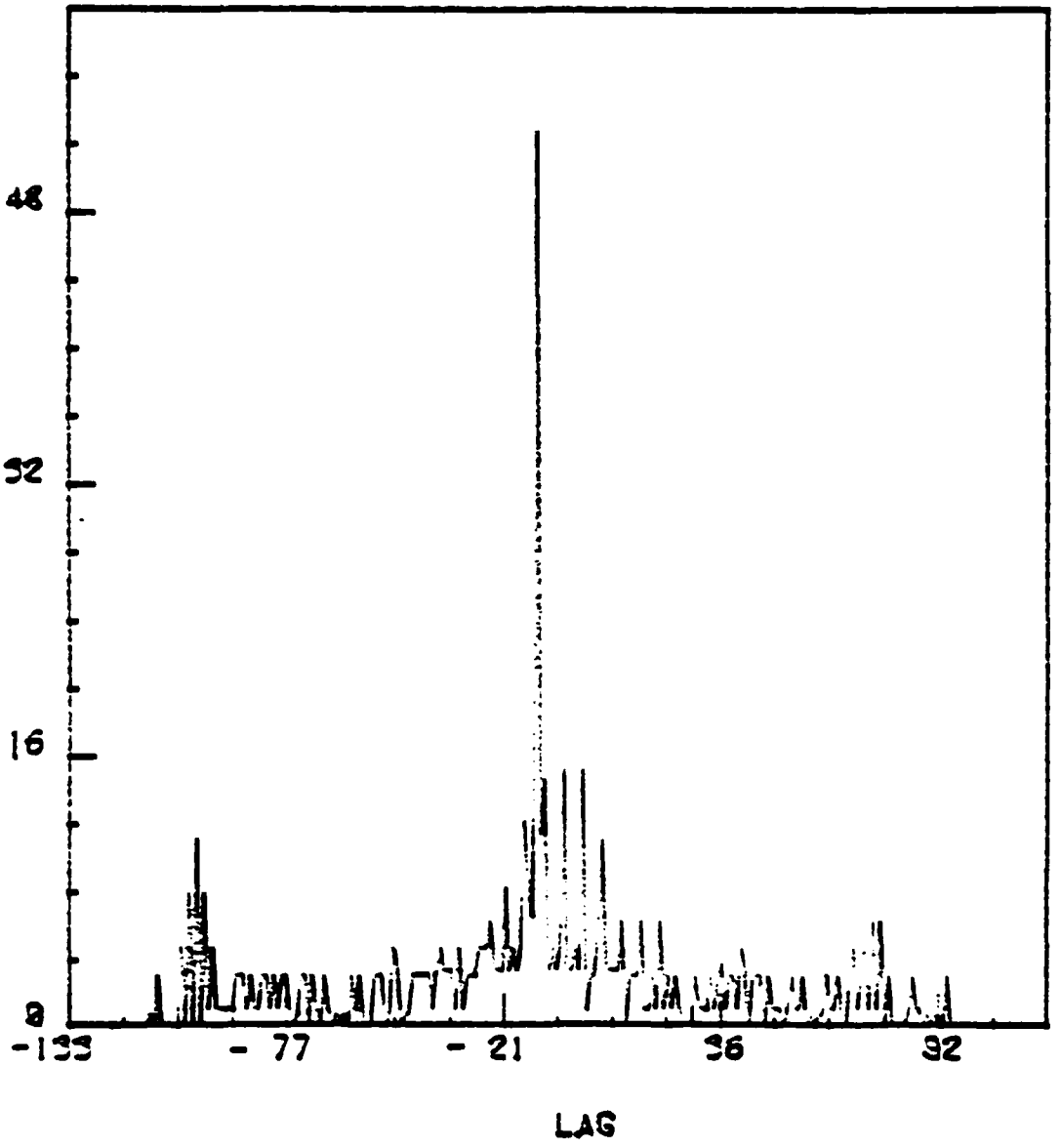


Figure 5-8. $M(p_0,p_1,k)$ plotted vs. lag k for chorion proteins 292 and 18B. p_0=.20, p_1=.90.

APPENDIX 1.

This appendix derives the GLLR test statistic for the hypothesis that within the string of independent binary variables $Z_1, Z_2, \ldots, Z_N$ there exists a substring $Z_i, Z_{i+1}, \ldots Z_j$ such that the probability of success within the substring is larger than that outside it. The GLLR test statistic above is related to the test statistic of equation (5-7).

Let

$$Z_1, Z_2, \ldots, Z_{i-1}, Z_i, \ldots, Z_j, Z_{j+1}, \ldots Z_N \tag{A-1}$$

be a string of independent binary variables. We shall refer to 1's and 0's as successes and failures. Suppose that the success probability for the substring

$$Z_1, Z_2, \ldots, Z_{i-1}, Z_{j+1}, \ldots, Z_N \tag{A-2}$$

is p_0 and that for

$$Z_i, Z_{i+1}, \ldots, Z_j \tag{A-3}$$

is p.

Let $0<p_0<p_1<1$. We are interested in testing the hypothesis

$$H_0: p=p_0 \text{ vs. } H_A: p \geq p_1. \tag{A-4}$$

If S_1 and S_0 are the numbers of successes and failures for the substring in (A-3) and T_1 and T_0 are the number of successes and failures in the substring in (A-2), under H_A,

$$\Pr(T_0=t_0, T_1=t_1, S_0=s_0, S_1=s_1)$$

$$= \begin{bmatrix} N-(j-i+1) \\ t_1 \end{bmatrix} p_0^{t_1} (1-p_0)^{t_0} \begin{bmatrix} j-i+1 \\ s_1 \end{bmatrix} p^{s_1} (1-p)^{s_0},$$

and under H_0,

$$\Pr(T_1+S_1=t_1+s_1)= \begin{bmatrix} N \\ s_1+t_1 \end{bmatrix} p_0^{s_1+t_1}(1-p_0)^{s_0+t_0}.$$

Hence the GLR for the hypothesis (A-4) is:

$$\frac{\sup_{i<j \quad p\geq p_1} \begin{bmatrix} N-(j-i+1) \\ t_1 \end{bmatrix} \begin{bmatrix} j-i+1 \\ s_1 \end{bmatrix} p^{s_1}(1-p)^{s_0} p_0^{t_1}(1-p_0)^{t_0}}{\begin{bmatrix} N \\ s_1+t_1 \end{bmatrix} p_0^{s_1+t_1}(1-p_0)^{s_0+t_0}}$$

$$= \frac{\sup_{i<j \quad p\geq p_1} \begin{bmatrix} N-(j-i+1) \\ t_1 \end{bmatrix} \begin{bmatrix} j-i+1 \\ s_1 \end{bmatrix} (p/p_0)^{s_1}((1-p)/(1-p_0))^{s_0}}{\begin{bmatrix} N \\ s_1+t_1 \end{bmatrix}}$$

$$= \max_{i<j} \frac{\begin{bmatrix} N-(j-i+1) \\ t_1 \end{bmatrix} \begin{bmatrix} j-i+1 \\ s_1 \end{bmatrix}}{\begin{bmatrix} N \\ s_1+t_1 \end{bmatrix}} \cdot \sup_{p\geq p_1} (p/p_0)^{s_1} ((1-p)/(1-p_0))^{s_0}.$$

Therefore the GLLR test statistic for the hypothesis that for some substring of $Z_1,\ldots,Z_N$ the success probability is not smaller than p_1 is

$$= \max_{i<j} \log \frac{\begin{bmatrix} N-(j-i+1) \\ t_1 \end{bmatrix} \begin{bmatrix} j-i+1 \\ s_1 \end{bmatrix}}{\begin{bmatrix} N \\ s_1+t_1 \end{bmatrix}} + L_{ij}(p_0,p_1),$$

for $L_{ij}(p_0,p_1)$ defined in equation (5-6). $M(p_0,p_1)$, the test statistic of chapter 5, neglects the first term and equals

$$\max_{i<j} L_{ij}(p_0,p_1).$$

BIBLIOGRAPHY

1. Bickel P.J. and Doksum K.A. (1977). Mathematical Statistics. Holden-Day, San Francisco.
2. Chernoff H. (1954). "On the Distribution of the Likelihood Ratio." Ann. Math. Stat. 28, 573-578.
3. Chung, K.L. (1974). A Course in Probability Theory. 2nd edition, Academic Press, New York.
4. Chvatal V. and Sankoff D. (1975). "Longest Common Subsequences of Two Random Sequences." J. Appl. Prob. 12, 306-315.
5. Deken J.G.(1976). "On Records: Scheduled Maxima Sequences and Longest Common Subsequence." Technical report No 91, Department of Statistics, Stanford University.
6. Dayhoff, M.O. (1972). Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Washington D.C.
7. Gibbs A.J. and McIntyre G.A. (1970). "The Diagram, a Method of Comparing Sequences. Its Use with Amino Acid and Nucleotide Sequences." Eur. J. Biochemistry 16, 1-11.
8. Hood, L.E., Wilson J.H. and Wood W.B. (1975). Molecular Biology of Eucariotic Cells. Benjamin/Cummings, Menlo Park, Calif.
9. Lehmann, E.L. (1975). Nonparametrics. Holden-Day, San Francisco, Calif.
10. Mahan, B.H. (1969). University Chemistry. Addison-Wesley, Reading, Mass.

11. Needleman S.B. and Wunch C.D.(1970). "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." J.Mol Biol. 48, 443-453.

12. Rao C.R. (1973). Linear Statistical Inference and Its Applications. 2nd edition, Wiley, New York.

13. Sankoff D. (1972) "Matching Sequences under Deletion/Insertion constraints." Proc. Nat. Acad. Sci. USA vol 69, No 1, 4-6

14. Steele M.J. (1980). "Long Common Subsequences and the Proximity of Two Random Strings." Technical report, Department of Statistics, Stanford University.

15. Watson, J.D. (1975). The Molecular Biology of the Gene. Benjamin, Menlo Park, Calif.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 36	2. GOVT ACCESSION NO. AD-A153 605	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Towards A Statistical Analysis of Genetic Sequences Data With Particular Reference To Protein Sequences		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Spyros P. Arsenis		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0555
9. PERFORMING ORGANIZATION NAME AND ADDRESS Statistics Center Massachusetts Institute of Technology Cambridge, Massachusetts 02139		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS (NR-609-001)
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics and Probability Code 436 Arlington, Virginia 22217		12. REPORT DATE March 1985 13. NUMBER OF PAGES 118
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

"Genetic Sequences, DNA, Matrix Smear, Character Matrix Graphics,

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See reverse side.

DD FORM 1473 EDITION OF NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT

This report develops a variety of character matrices as graphical tools for the visual examination of genetic sequences and in particular protein sequences. The NNC, PNC, BNC1, BNC2 and BNC3 matrices are designed to filter noise without severely suppressing signals in the CC matrix. The Matrix Smear of a character matrix is introduced as a measure of signals and noise in the matrix. The asymptotic distribution of the smears of the CC and NNC matrices are derived under the independence model. The asymptotic result is used in conjunction with exact confidence intervals from diagonal smears to automate partially the visual examination of character matrices. A generalized likelihood ratio procedure is developed to automate fully the detection of signals in two protein sequences. A simulation study has proven the procedure to be powerful and robust in detecting signals of success probability .90 and length 9 implanted within noisy binary strings of length 291 characters and success probability .15.

END

FILMED

6-85

DTIC